Multimodal OXYmorons: A Comprehensive Introduction and Computational Analysis Using a Dataset of Oxymoronic Memes in Italian and Spanish

Eliana Di Palma[1*], Giulia Rizzi[2,3*], Francesca Masini[4], Paolo Rosso[5], Elisabetta Fersini[2]

[1]University of Turin
[2]University of Milano-Bicocca
[3]Universitat Politècnica de València
[4]University of Bologna
[5]Universitat Politècnica de València
ValgrAI Valencian Graduate School and Research Network of Artificial Intelligence

This article introduces the concept of multimodal oxymorons. Multimodal oxymorons extend the traditional oxymoron theory by constructing and communicating meaning through the interplay of multiple modalities (such as visual and textual) rather than relying solely on language. We argue that multimodal oxymorons are central mechanisms of meaning-making in contemporary communication, as evidenced by the use of memes as an example.

While textual oxymorons have long been the subject of analysis in order to ascertain their role in shaping thought and meaning, multimodal oxymorons demonstrate how human cognitive process transcends linguistic boundaries, integrating different modalities (e.g., visual) in order to convey complex ideas. To encourage further study, we present a curated multilingual dataset of Multimodal OXYmoron (MOXY), which can be used as a foundation for further analysis and experimentation. Furthermore, we propose a methodical approach for the identification of multimodal oxymorons along with a pipeline for automated generation. Through illustrative examples and a detailed methodology, this work establishes a comprehensive framework for understanding, identifying, and generating multimodal oxymorons, paving the way for advancements in computational linguistics, artificial intelligence, and figurative language studies.

∗ E. Di Palma and G. Rizzi are co-first authors.

Action Editor: Wei Gao. Submission received: 05 February 2025; revised version received: 05 November 2025; accepted for publication: 23 November 2025.

https://doi.org/10.1162/COLI.a.586

1. Introduction

In the rapidly evolving domain of multimodal studies that encompass both linguistic and computational fields, considerable emphasis has been placed on the exploration of figurative language.

Multimodal figurative language refers to the use of figurative language (such as metaphors and similes) and idioms across multiple modes of communication, including text and images (Dancygier and Sweetser 2014; Forceville and Urios-Aparisi 2009; Veale 2008). Understanding multimodal figurative language is an important challenge in artificial intelligence, as it requires the integration of vision, language, commonsense, and cultural knowledge. Research in this area has led to the development of datasets and benchmarks for multimodal understanding of figurative language, with the aim of driving the development of models that can better understand and interpret figurative language across different modalities (Yosef, Bitton, and Shahaf 2023). This concept has been explored in various contexts, such as social media, advertising, and news, where figurative language is often conveyed through both text and images (Tomás et al. 2022).

Despite the extensive discourse and research dedicated to metaphors and idioms in general, there remains a notable gap in the investigation of the phenomenon of oxymorons (or *oxymora*). This discrepancy highlights the need for further inquiry into less-explored facets of figurative language (Athanasiadou 2024).

To fill this gap, the article introduces the concept of multimodal oxymorons, which expand traditional oxymoron theory by generating and conveying meaning through the interaction of multiple modalities, such as text and image, rather than relying solely on verbal language. It argues that multimodal oxymorons serve as key mechanisms of meaning-making in contemporary communication, with memes serving as a prominent illustrative example (see Figure 1).

Figure 1
An example of a multimodal oxymoron in a meme where the text reads 'the happiness that distinguishes me', while the accompanying image, specifically the woman's face from Edgar Degas' *L'absinthe*, visually conveys the opposite sentiment.

The article contributes as follows:

- RQ1 - What characterizes multimodal oxymorons?
- RQ2 - Do visual elements in multimodal oxymorons influence the variability and multiplicity of perceived semantic contrasts compared to textual oxymorons?
- RQ3 - To what extent are existing state-of-the-art models capable of accurately identifying and analyzing multimodal oxymorons, and what challenges or limitations do they face?
- RQ4 - Is it possible to define an effective pipeline for the automatic creation of a dataset of multimodal oxymorons, generating both linguistic and visual elements?

The rest of the article is organized as follows: In Section 2, existing literature on oxymorons is presented, also overviewing multimodal expressions of figurative language. Multimodal oxymorons are then introduced in Section 3, highlighting their commonalities and differences with both textual and visual ones; answering therefore RQ1. A comprehensive dataset of multimodal oxymorons is presented in Section 4, delineating the procedure adopted both for collecting the data and for the labeling phase. A particular focus has been given to the oxymorons' perception, both providing examples and characterizing the interpretation process of multimodal oxymorons (answering RQ2). In Section 5, a pipeline for multimodal oxymoron identification is proposed. Two different approaches have been investigated, focusing on Vision-Language models and on Large Language Models. Specifically, we evaluated two state-of-the-art vision-language models: `CLIP-ViT-B-32-multilingual-v1` (Radford et al. 2021; Reimers and Gurevych 2019) and `mBLIP-mt0-xl` (Geigle et al. 2024), as well as ChatGPT-4o-mini. The results from these experiments contribute to addressing RQ3. Additionally, focusing on RQ4, a procedure for the automatic generation of multimodal oxymorons is designed and evaluated in Section 6. Finally, in Section 7 conclusions are drawn.

2. Related Works

According to Gibbs and Kearney (1994), contemporary investigations within cognitive linguistics have underscored the significance of tropes as cognitive tools, playing an important role in everyday linguistic interactions and constituting fundamental components of everyday cognitive processes. These rhetorical tools help us make sense of contradictory situations, allowing us to comment on them using figurative language. For instance, oxymorons, like *bittersweet*, reflect our complex understanding of life's contradictions. In simpler terms, tropes help us express and understand contradictions in language and thought.

The following section presents an overview of existing literature, focusing on oxymorons and multimodal expressions of figurative language. In the first part, the phenomenon of oxymorons is compared with other phenomena concerning contradictory and/or contrasting elements, such as paradox, antithesis, irony, and antonymy, in order to evaluate their similarities and differences, especially in the multimodal domain.

2.1 Oxymorons

Oxymorons reflect a way of conceptualizing incongruous events by their peculiar mechanism of creating meaning by combining two conflicting terms in a forced manner (Gibbs and Kearney 1994).

Oxymorons share characteristics with other rhetorical figures such as paradox and antithesis, as all three involve the juxtaposition of contradictory elements. As a rhetorical device, an oxymoron can be understood as a condensed paradox, while a paradox may be seen as an expanded oxymoron (Ruiz de Mendoza Ibáñez 2020; Ruiz de Mendoza Ibáñez 2012; Ruiz 2009; Flayih 2009). In cognitive linguistics, a paradox is defined as a seemingly self-contradictory statement that emerges from an internal conflict between predefined interpretations of predications within a single utterance (e.g., *Freedom is slavery*). In contrast, **an oxymoron occurs when contrasting properties are attributed to the same entity** (e.g., *That was a* ***peaceful war***).

Antithesis, by comparison, is a rhetorical structure that places contrasting terms in parallel or balanced expressions (e.g., Plato's Republic: *justice consists in doing good to friends, evil to enemies*) (Fahnestock 1999). Tseronis and Forceville propose a definition of visual and multimodal antithesis (Tseronis and Forceville 2017). Briefly, they propose that, for a configuration to be classified as a visual or multimodal antithesis, it must meet the following three criteria: (1) it should involve two states of affairs, or entities, that are recognized or implied as opposites; (2) these opposites should be presented in a parallel structure that emphasizes their differences; and (3) the configuration should bring attention to the diametrically opposed viewpoints, ideas, or interests associated with the two entities or states of affairs within the given context. The main difference between antithesis and oxymoron, and consequently between multimodal antithesis and multimodal oxymoron, is that the former is based on some kind of parallel structure that invites comparison (as illustrated in Figure 2).

Moreover, oxymorons share with irony the use of oppositions to create humor. Irony is a multifaceted and complex phenomenon extensively studied in linguistics and philosophy. While the term can encompass a broader sense of humor or the humorous

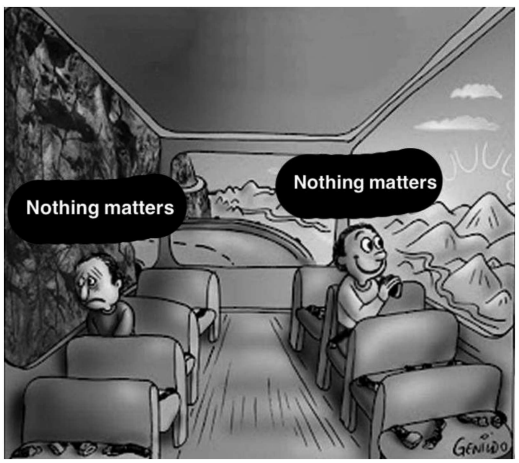


Figure 2
An example of multimodal antithesis with a parallel structure based on the meme Two guys on a bus (image created by Genildo Ronchi).

effect of a statement, verbal irony possesses distinct characteristics: such as the "echoing mention", the echoing of an attributed thought (e.g., belief, intention, or norm) while expressing a mocking, skeptical, or critical attitude. In a broader sense, irony may employ rhetorical figures like hyperbole, understatement, simile, metaphor, litotes, and insinuation to create humor or achieve rhetorical effects (Dynel 2016). This broader understanding extends to the perception of incongruity in memes. As Yus (2023) suggests, humor often arises from the process of resolving incongruity, such as in the case of oxymorons. In this study, we consider irony in its broadest sense, akin to a sense of humor when perceiving memes.

Finally, oxymorons are based on antonymy, a semantic relationship between words with opposite meanings. Essentially, antonyms are pairs of words that convey contrasting meanings.

Antonymic pairs are often used in texts and in many proverbs and idioms to achieve rhetorical effects (e.g., *a friend of everyone is a friend of no one*), but, more importantly for our current purposes, they are one of the ingredients of figures such as paradox, antithesis, irony, and, above all, oxymorons.

Textual oxymorons are juxtapositions of antonymic terms in the same phrase. Previous work on oxymorons relied precisely on antonymic pairs. La Pietra and Masini (2020), for instance, started from 17 antonyms present in Jones (2003) to build their dataset of Italian oxymorons. Their antonym list is itself a refined subset of Jones's (2003) English antonym list, which originally contained 56 pairs. For the Italian study, only noun-based antonyms were retained, excluding verb-based pairs such as *to give – to receive*. Their corpus-based approach using this list resulted in 377 oxymorons.

Similarly, Bolognesi et al. (2024) started from a list of adjectives drawn from two major Italian lexical resources, Tullio De Mauro's *Vocabolario di Base* (1980) and its updated version (2016) (De Mauro 2016). They identified 945 high-frequency adjectives (e.g., "*abile*" skillful, "*felice*" happy) and derived morphologically related nouns (e.g., "*abilità*" skill, "*felicità*" happiness, "*positività*" positivity), ultimately producing a dataset of 207 unique oxymorons.

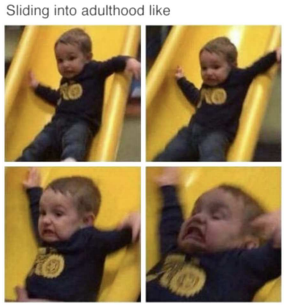
In the present work, as we will see, the pairs of antonyms are used both in the construction of the dataset of multimodal oxymorons (Section 4) and in the recognition of the oxymoron from the oxymoron meme (Section 5).

2.2 Multimodal Figurative Language

As previously mentioned, metaphors are one of the most investigated phenomena in studies on figurative language (Veale, Shutova, and Klebanov 2016), and the same holds for multimodal metaphors (Zhang et al. 2021; Jahameh and Zibin 2023; Zhong, Wen, and Chen 2023). A **multimodal metaphor** is a metaphor that draws on two or more modes/modalities to activate mapping between the tenor and the vehicle (Richard 1936), or target domain and source domain (Lakoff and Johnson 1980). Multimodal metaphors are defined by Forceville (2006) as metaphors wherein the target and source are predominantly or exclusively represented in different modes. The distinct modes, namely, verbal and visual, are posited by Kress and Van Leeuwen (2020) to possess their unique possibilities and limitations of meaning. In a multimodal context, meaning arises from the interplay of these diverse modes within their environmental context.

In Scott's work (2021), memes are considered as multimodal metaphors. In her analysis, Scott associates the labeling of memes with the creation of the multimodal metaphor. Each act of labeling creates a metaphorical relationship between the object or person in the text and the object or person described in the label (see the image

Table 1
Examples of multimodal figures by Lou (2017) and Benammar (2023).

Multimodal Figures	
	
Multimodal simile	*Multimodal metaphor*

on the right in Table 1). The incorporation of the text acts as an anchor, as discussed by Barthes (1977), guiding the viewer's interpretation of the image. Moreover, the text is not interpretable on its own and relies on the contextual support provided by the image. Both elements, image and text, often interact synergistically to shape the overall interpretation.

In addition to analyzing multimodal metaphors in the form of memes, Benamar also proposes a qualitative analysis of memes identifying multimodal similes and metonyms (Benammar 2023).

In Internet memes, similes rely on the verbal form, i.e., the expression *like* would be present within the textual part of the expression (see first image in Table 1). Lou (2017) broadens the definition of multimodal simile by considering cases in which, although not expressed verbatim, there is an explicit comparison between the situation described in the text and the image.

The corpus of research on multimodal figurative language, as with the broader field of figurative language studies, exhibits a pronounced emphasis on the analysis of metaphors and metonymies. In fact, although figurative language has been the subject of extensive study by researchers in a range of disciplines, their attention has been predominantly focused on the analysis of metaphors. This implies a significant research gap concerning other figures of speech, particularly oxymorons, especially in the multimodal domain.

This article introduces a novel concept that has yet to be systematically examined by scholars: the **multimodal oxymoron**.

3. Multimodal Oxymorons

In this section the concept of multimodal oxymorons is introduced in order to provide a definition highlighting the difference among unimodal and multimodal oxymorons and focusing on further distinctions.

3.1 Defining Multimodal Oxymorons

We define multimodal oxymoron as follows:

> *The juxtaposition of opposing or contrasting concepts within a single representation. For an instance to qualify as an oxymoron, the two juxtaposed concepts must be relevant to the same entity at a given time. Furthermore, to qualify it as multimodal, the oxymoron must manifest in a multimodal representation partially in one modality (e.g., visual imagery) and partially in another modality (e.g., textual form).*

In order to refer to the multimodal oxymoronic content, we will use, in this context, the pair of antonyms expressed in the two modalities (e.g., *maturity - immaturity*, *beauty - ugliness*), simply calling it "multimodal oxymoron". An overall overview of the different kinds of oxymorons, along with their characteristics, is summarized in Table 2.

In the realm of multimodal communication, concepts can seamlessly combine to express oxymorons. However, crucially, the presence of multiple modalities in the expression of an oxymoron does not imply that the oxymoron is multimodal. The

Table 2
Comparison of textual, visual, and multimodal oxymorons.

Aspect	Textual Oxymoron	Visual Oxymoron	Multimodal Oxymoron
Definition	A juxtaposition of opposing words.	A juxtaposition of visually contradictory elements.	A juxtaposition of opposing concepts expressed in different modes (e.g., one concept in the text and one in the image).
Medium	Text.	Images, designs, or artworks.	A combination of two media (e.g., image and text).
Expression	Semantic contrasts in words.	Visual contrasts.	Conceptual contrasts.
Interpretation	Relies on cognitive understanding of language.	Relies on visual perception and symbolic interpretation.	Relies on cognitive understanding of language and on visual perception and symbolic interpretation.
Purpose	To create humor, highlight contrasts, increase expressive intensity, or create new shades of meaning.	To evoke curiosity, tension, or deeper reflection through contrasting visuals.	To create humor, highlight contrasts, evoke curiosity, tension, or deeper reflection.
Processing	Sequential and processed linearly (read word by word).	Instantaneous and processed as a whole (viewed at once).	Sequential.
Example	‘‘Burning ice’’		

Table 3
Unimodal oxymorons expressed within a multimodal representation. The English translation of the original Italian text (represented within the image) is also reported.

Unimodal Oxymoron (Text)	Unimodal Oxymoron (Image)
	
The giraffe who suffers from vertigo: the queen of low heights.	*Trying to keep my anger under control.*

oxymoron might, in fact, be expressed in an unimodal way, namely, through only one of the modalities that form a multimodal element. In fact, if the information expressed in a multimodal content by a single mode alone is enough to identify the two contrasting concepts, the oxymoron counts as unimodal.

This phenomenon is particularly evident in the context of Internet memes, where oxymorons may be exclusively encapsulated either in the textual or in the visual content. Table 3 reports two examples of memes in which unimodal oxymorons (textual and visual, respectively) are expressed within a multimodal element.

As expressed in the definition 3.1, representing contrastive concepts in different modalities is a necessary but non-sufficient condition for multimodal oxymorons. Table 4 reports an example of a true multimodal oxymoron in meme versus three non-oxymorons. The first meme represents a multimodal oxymoron because the two contrastive concepts (*happy-sad*) are expressed within a single representation but through

Table 4
Oxymoron analysis. The English translation of the original Italian text (represented within the image) is also reported.

Multimodal Oxymoron	Non-oxymoron	Non-oxymoron	Non-oxymoron
			
I'm happy to go to the date.	*I'm happy to go to the date.*	*"I'm happy to go to the date". As soon as she sees me:*	*"I'm happy to go to the date". Me after the date:*

Table 5
Different contrastive concepts derived from a multimodal oxymoron. The English translation of the original Italian text (represented within the image) is also reported.

Multimodal Oxymoron	Contrastive Concepts
«Moderno appartamento in affitto» L'appartamento:  *"Modern apartment for rent"* *The apartment:*	• Moderno - antico [*Modern - ancient*] • Moderno - classico [*Modern - classic*] • Moderno - desueto [*Modern - obsolete*] • Moderno - passato [*Modern - past*] • Moderno - antiquato [*Modern - old-fashioned*] • Nuovo - vecchio [*New - old*]

two different modalities (*happy* as text, *sad* as image), referring to the same entity at a given time. The other memes reported, instead, are non-oxymorons since they do not meet all requirements. In the second meme in Table 4, both the image and the text report the same concepts (*happy-happy*) and not contrastive ones. In the third and in the last meme contrastive concepts are expressed in different modalities, but they do not refer to the same entity (third meme) or to the same time (fourth meme).

A peculiar aspect of multimodal oxymorons refers to the link among the contrastive concepts that characterize them. In the case of unimodal textual oxymorons, the opposing terms are stated within the text, leaving minimal space for subjectivity in the conceptualization of the elements in opposition, as the relationship is essentially explicit. The link among concepts is, therefore, *one-to-one*.

For visual oxymorons, instead, the subjectivity in the perception leads to a multiplicity of contrasting concepts, which results in a *many-to-many* relation.

In the case of multimodal oxymorons, on the one hand, the textual component leaves minimal space for subjectivity; while on the other hand, the varied interpretations and perceptions derived from the subjective nature of the visual content result in a multiplicity of contrasting concepts. This results in a one-to-many relationship between the overarching concept identified in the text and the divergent elements extracted by the image by different annotators. Table 5 provides an example of a multimodal oxymoron along with different combinations of concepts that can be identified[1].

In this specific example, the common element is the conceptualization of "moderno" (modern), which can be easily identified within the text. Conversely, the multiplicity of contrasting concepts ("antico", "classico", "desueto", etc.) reflects varied interpretations and perceptions derived from the subjective nature of the visual content.

While the concepts of "moderno" (modern) and "nuovo" (new) can be linguistically grouped as part of a broader concept linked to time, contemporaneity, and relevance,

1 The reported contrastive concepts have been collected involving non-expert annotators and have been linguistically validated. The contrasting concepts reported here are examples of ways of reading the oxymoron and are not intended to be an exhaustive list. Different perceptions may lead to other equally valid contrasting concepts.

since they overlap conceptually in representing what is current or of the now (although in a more cultural and stylistic dimension, or in a general and neutral one, respectively), the concepts derived from the image emphasize different aspects.

Such concepts—"vecchio", "antico", "classico", "desueto", "passato", and "antiquato"—all connect to the general concepts of age, the past, and obsolescence, but each emphasizes different nuances. Ranging from general oldness ("vecchio"), timelessness, and enduring quality ("classico") to obsolescence or disuse ("desueto"). While on the one hand, the concept conveyed within the text, although still subjected to interpretation, can be reconducted to a single concept, in this case, "modern"; on the other hand, the image arises in the annotators' multiple perceptions, linked with different concepts, that coexist at the same time (when looking at the image a user might perceive a classic, old-fashioned, obsolete, and old apartment). Among the perceptions raised by the image, some concepts might be contrastive to the ones raised by the text.

3.2 Further Categorization

In Shen's study (1987), oxymorons are categorized into "poetic" and "non-poetic", based on their presence in poetic corpora. Poetic oxymorons are further classified into "direct", "indirect", and "metaphorical". Direct oxymorons feature terms with a direct antonymous relation, like *living death*, while indirect oxymorons, exemplified by *sweet sorrow*, involve one term representing the hyponym of the other term's antonym (sorrow is a type of bitter entity). The latter are frequent in poetry and deemed the most poetic and understandable. Finally, metaphorical oxymorons contain terms that are not strictly speaking antonymic but belong to different conceptual domains (hence the contrast), as seen in *the silence goes*, where the (higher level) feature "+movement" referring to *goes* is not shared by *silence*.

Concerning multimodal oxymorons, within the confines of the given definition, distinctions emerge between oxymorons characterized by varying degrees of directness.

Similar to textual oxymorons, we can identify **indirect multimodal oxymoron**, which involves a more elaborate use of contradiction. The meaning of an indirect multimodal oxymoron is not immediately evident, as it requires a deeper understanding of the context surrounding one or both elements. In the case of memes, this could affect the textual component, the image, or both. Consequently, an indirect multimodal oxymoron is defined by scenarios in which at least one component of the oxymoron is not presented explicitly and is not immediately recognizable.

A **direct multimodal oxymoron** is characterized by the explicit articulation of the antonymic dyad splitting into the two modalities. In the case of memes, one concept is explicitly conveyed within the textual content, coupled with an image that promptly conveys the opposing concept.

Table 6 reports examples of direct and indirect multimodal oxymorons. In particular, the first meme qualifies as a direct multimodal oxymoron since both the image and the text explicitly report the concepts that compose the oxymoron (*mature - immature*). The second and third memes in Table 6 represent indirect multimodal oxymorons due to the implicit representation of the concepts that compose it in the text and in the image, respectively. In particular, while the concept of "diverse" or "different" is easily perceived from the picture, considering the side-by-side pose of the two girls, which leads to a comparison of their different body types, the concept of "equality" is expressed indirectly through the idiom *essere due gocce d'acqua* (lit. to be two drops of water, meaning "to be identical"). In the third meme, on the other hand, the concept of

Table 6
Examples of direct and indirect multimodal oxymorons. The English translation of the original Italian text (represented within the image) is also reported.

Direct	Indirect (text)	Indirect (image)	Indirect (text and image)
			
The maturity that distinguishes me:	*We are two drops of water.*	*We have completed also this puzzle.*	*The house-and-church girl on Instagram...*

"complete" is explicitly expressed within the text, while the image represents pieces of a puzzle disposed in a way that resembles the image of the completed puzzle, but the puzzle is factually incomplete. Finally, the last meme represents an indirect multimodal oxymoron where both text and image convey the opposite concepts indirectly. In this case, the text represents the concept of "chastity" or "purity" through the idiomatic binomial *casa-e-chiesa* (lit. house-and-church, meaning "pious, churchy") and the image implicitly refers to the concept of "impurity" or "unchastity" induced by the pose of the depicted animal.

To answer **RQ1**, a definition of multimodal oxymoron is proposed and contextualized within the theoretical framework of textual oxymorons. The interrelation of multimodal oxymorons with other linguistic and visual elements in communication has been studied in order to provide a comprehensive analysis. A summary of the characteristics of multimodal oxymorons, highlighting their differences from textual and visual ones, is also provided in Table 2.

4. Experimental Dataset

In the present study, a multimodal oxymoron is considered representative when it exhibits the characteristics required by the provided definition (reported in Section 3.1). An instance is regarded as a multimodal oxymoron if there exists at least one conceivable interpretation aligning with the given definition. The multimodal oxymorons that were included in the dataset exhibit the required characteristics, albeit with varying degrees of efficacy. We can define efficacy in this context as the degree to which an oxymoron achieves its intended effect, as measured by the level of recognition and understanding of the oxymoron among the intended audience.

Consequently, multimodal oxymorons can be categorized along two dimensions: directness and efficacy.

4.1 Dataset and Annotation Process

To create a comprehensive meme dataset, we performed two main activities: (i) we downloaded content from different social networks, such as Facebook, X, Instagram, and Reddit, and (ii) we created synthetic data according to the multimodal oxymoron definition.

The construction of our dataset is primarily based on two prior studies on Italian oxymorons: La Pietra and Masini (2020), and Bolognesi et al. (2024). In line with La Pietra and Masini (2020), we used a core list of antonym pairs as keywords to guide both the Web scraping and the synthetic meme creation. In this case, our dataset is based on 25 antonym pairs selected from the studies mentioned above. The 25 antonym pairs represent conceptual oppositions that can be expressed using different lexical labels. Consequently, the dataset includes multiple variants of each oxymoron, expressed through different labels and images in both Italian and Spanish, ensuring a diverse and representative set of examples.

The collected memes have been labeled by two authors of this article—both experts in computational linguistics and figurative language, who will be referred to as domain experts throughout the remainder of the article[2] (duplicates have been previously removed). Among the labeled memes, we obtained both oxymoronic and non-oxymoronic memes. Some memes include textual elements reported in the image itself (e.g., page names, watermarks). Specifically, this concerns 35% of the dataset. The majority of these cases (28%) involve acronyms, watermarks, or source attributions embedded within the image. The remaining portion includes numerical data or formulas present in certain images (4.3%), and textual elements related to commercial products (4.3%). The different textual elements often co-occur within the same meme.

To further explore the complex interpretation of multimodal oxymorons, we broadened our investigation through crowdsourcing, incorporating additional labels to capture diverse perspectives.

If domain experts evaluated whether the necessary criteria for multimodal oxymorons were met within the memes, participant annotations captured how these oxymorons were perceived. Specifically, their responses indicated whether the memes were perceived as effective examples of multimodal oxymorons and what conceptual opposition they conveyed.

Importantly, participants were not asked to specify whether each concept of the perceived oxymoron originated from the textual or visual component of the meme, and therefore were not required to link individual labels to a specific modality. Instead, they were explicitly asked to consider the meme as a whole. Participants were therefore encouraged to interpret the oxymoronic opposition as emerging from the interaction between the modalities, rather than isolating the source of each term. Therefore, the dataset does not include explicit annotations indicating whether each term comes from the image or the text.

The items of the dataset were labeled using the crowdsourcing platform Prolific according to the following primary questions:[3]

- In your opinion, is this meme a multimodal oxymoron?
- If so, which pair of opposing concepts do you associate the meme with?

2 The evaluation provided by a domain expert was centered on the validation of the characteristics that define a multimodal oxymoron, namely, the spatial-temporal requirements.

3 The questions contained in the questionnaire were in the two languages that constitute the dataset: Italian and Spanish. For convenience, we report here the English translation of the questions. Mother-tongue Italian and Spanish speakers were selected for the annotation of the respective memes. The guidelines for participants and an extract of the labelling form are reported in Appendix 9.

Table 7
Dataset characteristics.

Subset	Non-oxymoron	Direct Oxymoron	Indirect Oxymoron		
			Text	Image	Text + Image
Italian	32	81	9	28	9
Spanish	28	87	11	11	3
Overall	60	168	20	39	12

The final **M**ulimodal **Oxy**morons dataset (**MOxy**) comprises a total of 300 memes in two languages, with 160 memes in Italian and 140 in Spanish.[4] The dataset includes, as "control" cases, one-fifth of the memes assessed as non-oxymorons by domain experts. This inclusion validated the annotators' evaluations' accuracy and consistency, laying the groundwork for further investigation into the complex dynamics of multimodal oxymoron perception. To ensure a rigorous and reliable analysis of the meme dataset, we organized the data into distinct online forms, each comprising 15 memes, with 5 of them designated as control memes (i.e., memes identified as non-oxymorons by domain experts).[5] For every form, we collected a total of 15 valid annotations from various annotators. Consequently, the final dataset encompasses, for each meme, the perspectives of 15 annotators. Such annotations are not combined into a single ground truth label. Instead, we intentionally preserve the individual annotations to reflect the inherent subjectivity of interpreting multimodal oxymorons. Table 7 summarizes the dataset characteristics.

The final dataset contains the following information for each meme:

- **Meme**: the meme itself in a jpg format;
- **Language**: the language of the meme (Italian or Spanish);
- **Expert evaluation**: the label introduced by experts to determine whether the meme contains a multimodal oxymoron;
- **Complexity evaluation**: the label introduced by experts to determine whether the multimodal oxymoron is direct or indirect;
- **Degree of efficacy**: the number of ordinary annotators that identified the multimodal oxymoron;
- **Perceived multimodal oxymorons**: the pair of opposite concepts expressed by the oxymoron as perceived by the recruited annotators and the corresponding distributions.

4 The dataset is publicly available for research purposes and can be accessed by filling out a request form, where users are asked to agree to the terms of use under the specified research license. To request MOxy: `https://forms.gle/3Hcxa9oh74kpf3HY8`.

5 To ensure the dataset's integrity, all annotations acquired from an annotator who did not correctly mark the control elements were discarded.

Table 8
Example of an entry in the dataset with the respective labels and metatada. Although not provided as part of the dataset, the English translation of the original Spanish text (represented within the image) is also reported.

Row image	Corresponding labels
"Es tan interesante. Por favor, cuéntame más"  *It's so interesting.* *Please, tell me more about it.*	• Language: Es • Expert evaluation: 1 • Complexity evaluation: Direct • Annotators evaluation: 15/15 • Perceived Multimodal Oxymorons: [((Interesante - aburrido), 13/15), ((interés - hartazgo), 1/15), ((Aburrido - divertido), 1/15)]

We report in Table 8 an example of a meme composed of a raw image and the corresponding labels available through a csv file.[6]

This dataset provides a useful resource for investigating how experts and laypeople understand multimodal oxymorons in memes.

The primary goal of the proposed dataset is to establish a replicable methodology for the systematic construction of oxymorons and to introduce and evaluate a phenomenon that was unexplored in both linguistic and computational literature. In contrast to large-scale Web-scraped datasets commonly used in NLP, our dataset is deliberately curated with a focus on precision rather than scale. This design choice aligns with the aims of our study, which centers on the careful identification and analysis of a specific and complex linguistic phenomenon. By prioritizing quality over quantity, we aim to minimize issues such as noise, redundancy, and labeling errors, and ensure the reliability and interpretability of our data.

4.2 Oxymoron Perception

As introduced in Section 3, subjectivity plays an important role within multimodal oxymorons. The visual component of the meme might, in fact, be perceived differently by the annotators and lead, therefore, to diverse perceptions and interpretations of the oxymoron. Such a difference in interpretation emerges at first in the nature of the multimodal oxymoron itself.

In a multimodal oxymoron, one part of the antonymic pair is associated with the text, while the different interpretation of the image introduces multiple possible opposite concepts.

Therefore, text-image oxymorons, as well as image-only, appear to be less constrained in their interpretation than text-only oxymorons.

6 Perceived Multimodal Oxymorons [ENG]: [((Interesting - boring), 13/15), ((Interesting - boredom), 1/15), ((Boring - fun), 1/15)].

Table 9
Different perceptions of a multimodal oxymoron. The English translation of the original Italian text (represented within the image) is also reported.

Multimodal Oxymoron	Perceived Oxymoron
 *...and a Coke Zero, to keep it light.*	• Verità illusoria [*Illusory truth*] • Verità - illusione [*Truth - Illusion*] • Magrezza - grassezza [*Thinness - fatness*] • Leggero - pesante [*Light - heavy*] • Contenuta ingordigia [*Contained greed*] • Dietetico - calorico [*Dietetic - caloric*] • Abbondanza - dieta [*Abundance - diet*] • Sano - spazzatura [*Healthy - garbage*]

In the meme reported in Table 9, for example, two different relationships are present:

Such findings and characteristics on the interplay among the concepts that define the oxymoron contribute to answering **RQ2**: While textual oxymorons effectively activate frames associated with opposing concepts, creating a direct contrast between terms like "old" and "modern" in the phrase "old modernity", multimodal oxymorons evoke multiple potential concepts simultaneously. The image component of a multimodal oxymoron can evoke a range of concepts, one of which is then contrasted with the concept conveyed by the text. The image component can evoke multiple concepts, leading to a variety of perceived contrasts depending on which elements of the image and text the viewer chooses to focus on. The juxtaposition of the shared element with a spectrum of contrasting concepts emphasizes the inherently subjective nature of visual stimuli.

This phenomenon aligns with findings from Cerini et al. (2022), who explored the associations between abstract concepts and visual stimuli. Their study demonstrated that even when images were selected to convey the same abstract concept, participants often associated them with different abstract terms. These results indicate that the semantics of a visual scene significantly influence the interpretation and preference for a given abstract concept. The variation in word-image associations, despite consistent conceptual intent, supports the idea that visual elements in multimodal communication evoke a diversity of meanings, shaped by contextual cues and individual perspectives. Thus, the interpretive multiplicity observed in multimodal oxymorons can be understood as an extension of this broader cognitive tendency toward variable grounding of abstract meaning in visual stimuli.

5. Multimodal Oxymoron Recognition

Two different approaches for the identification of multimodal oxymoron have been investigated, focusing on vision-language and on large language models, respectively. In addition, the performances of the proposed approaches are evaluated in relation to a random baseline.

5.1 Vision-Language Models

We evaluated two state-of-the-art vision-language models: `CLIP-ViT-B-32-multilingual-v1`[7] (Radford et al. 2021; Reimers and Gurevych 2019) and `mBLIP-mt0-xl`[8] (Geigle et al. 2024). Vision-language models use both visual and textual information to provide a comprehensive understanding of the input content. In particular, `CLIP-ViT-B-32-multilingual-v1` and `mBLIP-mt0-xl` are multilingual vision-language models designed for cross-modal tasks involving images and text in multiple languages.

`CLIP-ViT-B-32-multilingual-v1` combines the original CLIP model (Radford et al. 2021) image encoder with a multilingual text encoder (trained on over 50 languages) in a shared embedding space exploiting multilingual knowledge distillation. On the other hand, `mBLIP-mt0-xl` is a multilingual adaptation of BLIP-2 that adopts a ViT vision encoder, a Q-Former module, and the mT0-XL multilingual language model. mBLIP supports 96 languages and is trained via lightweight realignment of the English BLIP-2 model using multilingual instruction data.

Both models were fine-tuned for multimodal oxymorons detection using a 10-fold cross-validation strategy to ensure robustness and generalization of the results. The models were trained to jointly interpret textual and visual inputs and classify whether an input expresses a multimodal oxymoron or not.

To address the issue of class imbalance when evaluating vision-language models, we adopted Youden's J statistic (Youden 1950) to estimate the decision threshold. Using a fixed threshold of 0.5 caused the models to predict the majority class by default, thus failing to capture the true distribution of multimodal oxymorons. Computing the Youden J statistic on a validation set identifies a threshold that balances true positive and true negative rates more effectively. This adjustment significantly improves models' performances, which would otherwise underperform due to skewed predictions.

Table 10 illustrates the performance of vision-language models—`CLIP-ViT-B-32-multilingual-v1` and `mBLIP-mt0-xl`—across three datasets: the full dataset, the Italian subset, and the Spanish subset, for the detection of multimodal oxymorons. To compare the selected models, we measure Precision (P), Recall (R), and F-Measure (F1), distinguishing between the multimodal oxymorons (+) and the non-oxymorons one (−).

On the overall dataset, multilingual CLIP achieves the highest $F1^{+}$ score (0.40), indicating a better performance in detecting oxymorons compared to mBLIP (0.31). However, the overall F1-score remains close across all methods—multilingual CLIP (0.36), mBLIP (0.31). The proximity to the random baseline, which outperforms both visual-language approaches, further emphasizes the difficulty of the task. The same findings are likewise observable when considering the Italian and Spanish subsets individually.

7 `https://huggingface.co/sentence-transformers/clip-ViT-B-32-multilingual-v1`.
8 `https://huggingface.co/Gregor/mblip-mt0-xl`.

Table 10
Comparison of the different approaches for oxymoron detection. **Bold** denotes the best approach according to the F1-score, while underline represents the best approach according to the oxymoron label. (†) denotes that the model behave differently with respect to the random baseline, according to a McNemar's test ($p < 0.05$).

Approach	Oxymoron - On the Overall Dataset						
	Precision$^+$	*Recall*$^+$	*F1*$^+$	*Precision*$^-$	*Recall*$^-$	*F1*$^-$	*F1-score*
CLIP-ViT-B-32-multilingual-v1	0.82	0.26	0.40	0.21	0.77	0.33	0.36
mBLIP-mt0-xl	0.76	0.20	0.31	0.19	0.75	0.30	0.31†
Random	0.75	0.42	0.54	0.16	0.45	0.24	**0.39**

Approach	Oxymoron - Italian						
	Precision$^+$	*Recall*$^+$	*F1*$^+$	*Precision*$^-$	*Recall*$^-$	*F1*$^-$	*F1-score*
CLIP-ViT-B-32-multilingual-v1	0.89	0.26	0.40	0.23	0.88	0.36	**0.38**
mBLIP-mt0-xl	0.69	0.17	0.28	0.17	0.69	0.28	0.28†
Random	0.75	0.41	0.53	0.16	0.44	0.23	**0.38**

Approach	Oxymoron - Spanish						
	Precision$^+$	*Recall*$^+$	*F1*$^+$	*Precision*$^-$	*Recall*$^-$	*F1*$^-$	*F1-score*
CLIP-ViT-B-32-multilingual-v1	0.75	0.27	0.39	0.18	0.64	0.28	0.34
mBLIP-mt0-xl	0.83	0.22	0.35	0.21	0.82	0.33	0.34
Random	0.76	0.43	0.55	0.17	0.46	0.25	**0.40**

Additionally, the Nadeau and Bengio (N&B) (Nadeau and Bengio 1999, 2003) corrected paired t-test is used to assess whether the models' performances differ significantly (on the overall dataset) with respect to the random baseline. Unlike the standard paired t-test, the N&B test adjusts for the variability across the cross-validation folds, allowing for a more accurate comparison of model performance while accounting for the correlation introduced by the overlapping training sets across folds. In particular, we compared the F1-scores of the two models across three evaluation runs using the N&B corrected paired t-test, which assesses whether the mean difference in performance between the models is statistically significant ($p < 0.05$).

We also applied a McNemar's test to compare the actual predictions of the models on a per-sample basis. This non-parametric test evaluates whether two models differ significantly in the number of examples they classify correctly and incorrectly, providing a complementary perspective on performance differences beyond aggregate metrics.

As shown by the results, the performance of the vision-language models on this task is similar to the random baseline, highlighting the problem's inherent complexity and difficulty. The N&B corrected paired t-test on the model predictions revealed no statistically significant difference in overall macro F1-score between the models and the random baseline. This suggests that, on average, the models do not outperform random chance, and vice-versa. However, a McNemar's test comparing the prediction errors of mBLIP and the baseline indicated a statistically significant difference in their error

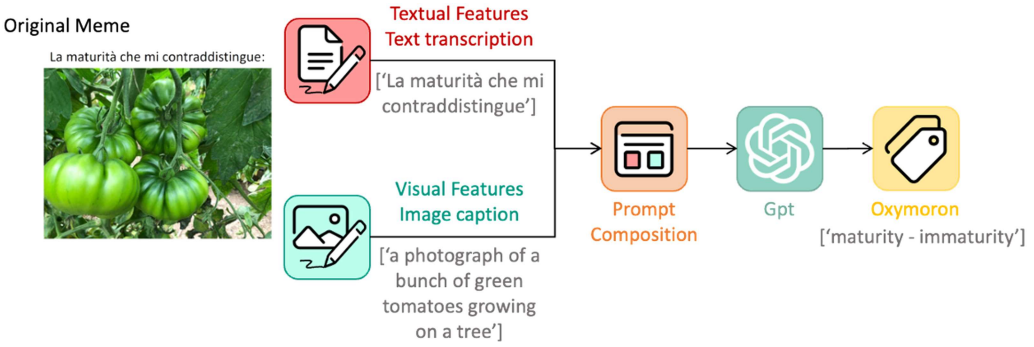


Figure 3
Schematic representation of the proposed pipeline for multimodal oxymoron identification.

distributions. This suggests that, while mBLIP may not achieve a statistically different F1-score, it exhibits different patterns of behavior and makes distinct types of errors.[9]

It is important to note that, although these models are fine-tuned for oxymoron identification, they do not explicitly detect the specific oxymoronic expression or the underlying antonymic relationship that composes it. For this reason, we further investigated large language models (LLMs), which are better suited to capture fine-grained semantic and syntactic relations within text.

5.2 Large Language Models

A pipeline for multimodal oxymoron identification has been defined. A key step of the proposed framework is the definition of a prompt, which consists of a definition of multimodal oxymorons, an example of a multimodal oxymoron meme, and an inference instance (i.e., meme to be predicted). The overall approach is summarized in Figure 3.

5.2.1 Meme's Components Elaboration. The first step of the pipeline aims at reconducting the whole meme to text by maintaining all the information that might be relevant for the identification of the oxymoron.

The two modalities that compose the meme (i.e., text and image) have been elaborated separately as follows:

- **Textual component**: The textual component of the meme has been represented through manual transcription of the text contained in the memes, as provided in the MOxy dataset.
- **Visual component**: The visual component has been reconducted to text via Image Captioning and Image Tagging operations, which generate a textual description and a list of image tags for each meme, respectively.

9 According to a McNemar's test, mBLIP does not exhibit different patterns of behavior with respect to the random baseline when considering the Spanish subset alone.

- **Image captioning**: In order to generate image captions, we adopted the Bootstrapping Language-Image Pre-training (BLIP) model (Li et al. 2022), a VLP framework with state-of-the-art performance on a wide range of downstream vision-language tasks, including understanding-based and generation-based tasks.[10]
- **Image tagging**: In order to generate image tags, we adopted the Recognize Anything Model (RAM) (Zhang et al. 2023). Unlike traditional models, RAM does not rely on manual annotations for training; instead, it leverages large-scale image-text pairs.[11]

By using image captioning and tagging, the oxymoron can be mapped into a unified textual modality, which allows for the identification of its constituent antonyms in a shared latent space while ensuring that the process remains replicable without incurring substantial computational costs.

5.2.2 Prompting. We investigated two prompting strategies: Zero-Shot and One-Shot. In both cases, prompts were meticulously crafted to include traditional key components commonly cited in prompt design literature (Choi and Chang 2025; Schulhoff et al. 2024): a clear definition of the novel concept (i.e., multimodal oxymoron), a task formulation, and a breakdown of the elements to consider. In the one-shot setting, this base structure was extended with a representative example to provide additional context and aid model understanding.

Other prompting strategies, i.e., direct and few-shot prompting, were excluded due to constraints specific to this task: The novelty of the multimodal oxymoron requires explicit guidance (which direct prompting lacks), and the limited number of representative oxymorons makes few-shot prompting impossible.

We also excluded traditional Chain-of-Thought (CoT). The rationale behind this choice lies in the characteristics of the approach itself. First of all, the CoT is linear and sequential, thus inducing a step-by-step reasoning process that is not suitable for the detection of multimodal oxymorons. Instead, a parallel comparison between the two modalities is required (e.g., the semantic content of the text vs. that of the image or description) to assess their semantic or symbolic dissonance (Zheng et al. 2025). The multimodal oxymoron is not a causal chain of thought; rather, it is a contrast among different modalities that should be compared simultaneously. Furthermore, CoT is designed for deductive or inductive reasoning, not for symbolic semantic conflicts. A CoT approach tends to force the logical coherence of the text, whereas the oxymoron resides precisely in dissonance, which the CoT might try to "solve" instead of highlighting it (Turpin et al. 2023).

In the proposed pipeline, the prompt serves as a comprehensive guide, encapsulating several crucial elements to facilitate the multimodal oxymorons' identification. The definition of the prompt incorporates the following key components:

10 We adopted BLIP obtained from the HuggingFace model repository (*Salesforce/blip-image-captioning-base*). The default parameters provided by the authors were utilized in our experiments.

11 We adopted pretrained RAM obtained from the official Github repository (`https://github.com/xinyu1205/recognize-anything?tab=readme-ov-file`). The default parameters provided by the authors were utilized in our experiments.

- **Definition of multimodal oxymoron**: The prompt elucidates the concept of a multimodal oxymoron, providing a clear and concise definition (see Section 3). This definition emphasizes the multimodal nature of the content.
- **Example of multimodal oxymoron meme**: To further illustrate the concept, the prompt gives an example of a multimodal oxymoron meme. This example serves as a reference point, providing a concrete representation of the desired outcome. It helps the model understand the intended humor and juxtaposition of concepts typical of multimodal oxymorons.
- **Task formulation**: The prompt explicitly outlines the task at hand, delineating the specific output that the model is expected to accomplish. In this case, the identification of the multimodal oxymoron is expected.
- **Description of the meme to be predicted**: A detailed description of the meme to be predicted is included in the prompt. This description encapsulates the text transcription of the meme and the image tags and/or caption. In particular, three different variations of the visual component within the prompt are investigated:
  - **Image tag-based**: The description of the visual component of the meme is provided via a list of tags.
  - **Image caption-based**: The visual component of the meme is represented through an image caption.
  - **Image tag- and caption-based**: Both image tags and image captions are encapsulated within the prompt.

By incorporating these elements into the prompt, as shown in Figure 4, we provide the model with a comprehensive understanding of the task, allowing it to navigate the complexities of multimodal oxymoron identification, ensuring that the generated output aligns with the desired characteristics outlined in the prompt.

Note that we encountered a multilingual scenario where the text transcription, in either Italian or Spanish, coexists with prompt instructions and image captions in English. In navigating this linguistic diversity, we decided to abstain from translating the transcriptions into English. The rationale behind this choice lies in the fact that translations may introduce variations in meaning and alter the selection of words, resulting in a modified representation of the memes, impacting the overall fidelity of the content and the oxymoron identification.

Acknowledging the intricacies involved in multilingual content composition, we opted for GPT (Generative Pre-trained Transformer),[12] a language model renowned for its capability to comprehend and reason with mixed-language prompts. GPT's proficiency in understanding and processing diverse linguistic inputs allowed us to maintain the authenticity of the original transcriptions while seamlessly integrating them with English prompts and captions. This strategic use of GPT ensured that the multilingual

12 We adopted ChatGPT 4o-mini. The model was accessed in April 2025; additional details are available at `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/`.

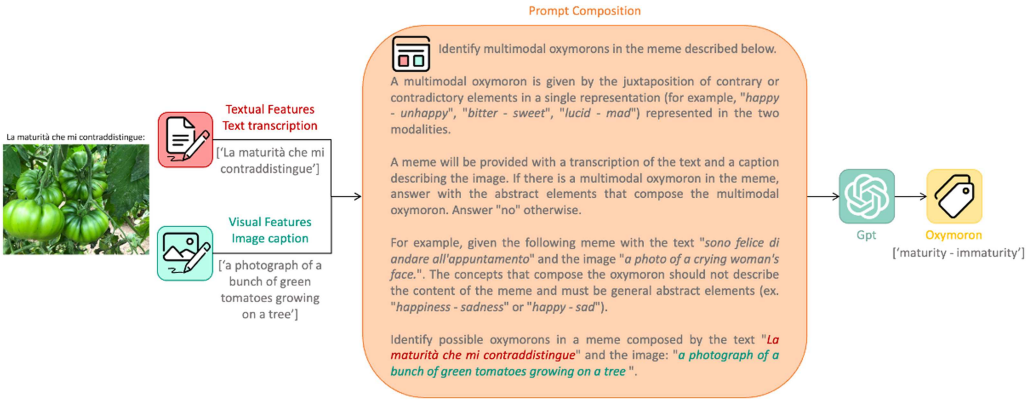


Figure 4
Schematic overview of the proposed pipeline for multimodal oxymoron recognition, exemplifying a prompt based on image captioning.

essence of the content was preserved, contributing to a more accurate and contextually relevant representation of memes within our study.

5.2.3 Results. In this study, we investigate the efficacy of three distinct approaches, namely, GPT_Tags, GPT_Capt, and GPT_Tags+Capt, considering both Zero-Shot and One-Shot prompting strategies, for the task of multimodal oxymoron detection. All the approaches have been evaluated on the proposed MOxy dataset, running multiple trials for each input stimulus to account for response variability, and aggregating the results (e.g., by averaging performance metrics across runs) to ensure robustness and reduce the influence of stochastic generation. In this section, achieved results are reported, also distinguishing between the two languages that compose the dataset. We measured Precision (P), Recall (R), and F-Measure (F), distinguishing between multimodal oxymoron (+) and non-oxymoron (−) labels and reporting also the Macro F1 (*F1-score*).

Table 11 summarizes the results achieved on the Multimodal Oxymorons Dataset, also considering both an overall estimation and the distinction between languages. Surprisingly, the achieved results highlight that the Zero-Shot approach based on image tags is the most-performing approach, achieving the highest overall F1-score of 0.53 on the overall dataset. In fact, a Zero-Shot approach based on tags, which, being decontextualized, is generally less semantically rich than captions, achieves the best performance in the identification task. Such a result is likely due to the concise and discriminative nature of the tags, which might help the model to focus on contrasting elements directly. Notably, the Zero-Shot approach based on captions and the One-Shot approach based on tags also demonstrate competitive performance.

These findings highlight the generally inferior performance of the tags and captions based approaches (i.e., Tags+Capt), as indicated by the lowest overall F1-score of 0.49 and 0.48 considering Zero- and One-Shot, respectively.

Overall, the achieved results confirm the ability of the model to handle multilingual prompts. All the proposed models are able to overcome the difficulties that may be encountered when dealing with multilingual prompts. All the identified oxymorons are generated in English, according to the prompt specification.

To evaluate model performance across three executions, we adopted paired t-tests to compare the average performance metrics (i.e., macro F1-score) across the same

Table 11
Comparison of the different approaches for oxymoron detection. **Bold** denotes the best approach according to the F1-score, while underline represents the best approach according to the oxymoron label. (*) denotes that the model outperforms the random baseline and obtains results that are statistically different according to a paired t-test ($p < 0.05$), while (†) denotes that the model behaves differently with respect to the random baseline, according to a McNemar's test ($p < 0.05$).

Oxymoron - On the Overall Dataset								
Approach	Input	*Precision*$^+$	*Recall*$^+$	*F1*$^+$	*Precision*$^-$	*Recall*$^-$	*F1*$^-$	*F1-score*
Zero-Shot	Tags	0.81	0.88	0.85	0.28	0.18	0.22	**0.53***†
	Capt	0.81	0.88	0.84	0.24	0.16	0.19	0.51†
	Tags+Capt	0.80	0.89	0.84	0.20	0.11	0.14	0.49†
One-Shot	Tags	0.81	0.88	0.84	0.25	0.15	0.19	0.51†
	Capt	0.80	0.87	0.83	0.17	0.10	0.12	0.48†
	Tags+Capt	0.80	0.88	0.84	0.18	0.11	0.13	0.48†
Random	–	0.80	0.50	0.61	0.20	0.49	0.28	0.45

Oxymoron - Italian								
Approach	Input	*Precision*$^+$	*Recall*$^+$	*F1*$^+$	*Precision*$^-$	*Recall*$^-$	*F1*$^-$	*F1-score*
Zero-Shot	Tags	0.82	0.90	0.86	0.33	0.19	0.24	0.55*†
	Capt	0.83	0.89	0.86	0.36	0.25	0.29	**0.58**†
	Tags+Capt	0.80	0.94	0.86	0.19	0.05	0.08	0.47†
One-Shot	Tags	0.80	0.91	0.85	0.17	0.07	0.10	0.48†
	Capt	0.80	0.89	0.84	0.19	0.10	0.13	0.49†
	Tags+Capt	0.79	0.93	0.86	0.07	0.02	0.03	0.44†
Random	–	0.80	0.49	0.61	0.20	0.51	0.29	0.45

Oxymoron - Spanish								
Approach	Input	*Precision*$^+$	*Recall*$^+$	*F1*$^+$	*Precision*$^-$	*Recall*$^-$	*F1*$^-$	*F1-score*
Zero-Shot	Tags	0.81	0.86	0.83	0.24	0.18	0.20	0.52*†
	Capt	0.78	0.86	0.82	0.08	0.05	0.06	0.44†
	Tags+Capt	0.80	0.83	0.82	0.21	0.18	0.19	0.51†
One-Shot	Tags	0.82	0.85	0.84	0.29	0.24	0.26	**0.55***†
	Capt	0.79	0.86	0.82	0.14	0.10	0.11	0.47†
	Tags+Capt	0.79	0.82	0.81	0.18	0.15	0.16	0.49†
Random	–	0.79	0.50	0.61	0.19	0.46	0.27	0.44

dataset. This allowed us to assess whether the observed performance differences were statistically significant under the assumption of normality.

A paired t-test on the model predictions revealed no statistically significant difference in overall macro F1-score between the majority of the models and the baseline. The best performing model (i.e., Zero-Shot based on image tags) achieves, according to the same statistical test, significantly better performances with respect to the random baseline, both considering the overall dataset and the language-defined subsets.

Furthermore, the McNemar test demonstrates that all the models behave significantly differently compared to the random baseline, despite achieving similar overall F1-score.

Table 12
Comparison of the different approaches for oxymoron detection on the overall dataset in terms of accuracy. **Bold** denotes the best approach for the corresponding oxymoron type.

Approach	Input	Oxymoron - Accuracy				Non-Oxymoron
		Direct	Indirect-Text	Indirect-Image	Indirect-Multi	
Zero-Shot	Tags	0.88	0.84	0.91	0.81	0.15
	Capt	0.88	0.83	0.89	0.83	**0.18**
	Tag+Capt	**0.89**	**0.90**	**0.92**	0.81	0.11
One-Shot	Tags	0.85	0.89	0.83	0.86	0.10
	Capt	0.86	0.89	0.89	**0.94**	**0.15**
	Tag+Capt	0.88	0.84	0.89	0.89	0.11

While the paired t-test reveals no statistically significant improvement in aggregate performance metrics, the McNemar results highlight that the models are learning some structured representations that influence their predictions in a non-random manner.

Additional analyses have been performed to assess the model's ability to identify direct and indirect multimodal oxymorons. Table 12 summarizes the achieved results considering different types of multimodal oxymorons. Model performance is reported in terms of accuracy. Achieved results highlight that the model based on both tags and captions (Zero-Shot Tag+Capt) demonstrates superior performance across various types of oxymorons, including direct oxymorons, indicating its effectiveness in multimodal oxymoron detection. The results show a general difficulty in recognizing the absence of multimodal oxymorons.

5.3 Multimodal Oxymoron Validation

While the previous analysis focuses on the abilities of state-of-the-art models to identify the presence of multimodal oxymorons in memes, additional analyses have been performed to evaluate the quality of the predicted oxymorons. The analysis presented above focused on the recognition of multimodal oxymorons as a binary classification task; the following analysis will instead focus on the validation of the concepts that make up the oxymorons, in order to assess not only the ability of the models to recognize multimodal oxymorons, but also the quality and validity of the contrastive concepts identified.

In particular, each predicted oxymoron has been validated by performing the following two steps:

- The labels of the concepts that compose the predicted oxymoron have been compared with the ones that compose the oxymorons gathered by the annotators, without considering the order in which they appear. If at least one match has been identified, the oxymoron is considered valid. This approach allowed us to account for linguistic variability in the results. For instance, *happiness-sadness* is considered a valid oxymoron even if the annotators have expressed it in the form *happy - sad*, *sad-happy*, or *sad happiness*.
- A second validation step was conducted by domain experts to manually evaluate the oxymoron pairs generated by the model in cases where no

corresponding examples were produced by the participants. This step aimed to capture potential interpretations beyond those provided by humans. Pairs were deemed valid if they represented meaningful opposites and were congruent with the image, and invalid if they lacked opposition or were incongruent.

The achieved results are summarized in Table 13.

The first column shows the percentage of multimodal oxymorons identified by the model. The second column shows the percentage of correctly predicted oxymorons, i.e., those cases where the model's prediction matches the crowdsourced antonym and/or is validated by experts.

While the Zero-Shot approach based on tags+captions appeared to be the most effective approach for identification, One-Shot prompting using captions yielded the best results for predicting actual oxymoronic concepts.

This finding aligns with our expectations, as the captions support deeper semantic interpretation and help the model to identify concept-level contradictions, also verifying the actual characteristics of multimodal oxymorons.

Focusing on the Zero-Shot approaches, the achieved results confirm the overall ability of the tags+caption-based model to predict multimodal oxymorons. This model has been shown to demonstrate a superior capacity in identifying oxymorons (89.17% identified oxymorons), and also exhibits a greater ability to generate valid oxymoronic expressions (more than 39% of valid oxymorons).

However, when evaluating the models' ability to generate valid multimodal oxymorons, the inclusion of an example in the prompt (i.e., adopting a One-Shot approach) leads to a substantial improvement in performance, with an increase of up to 24%. Within the One-Shot setting, the caption-based model emerges as the most effective, achieving a validity rate of 58% for the generated multimodal oxymorons. This model not only achieves competitive results in oxymoron identification but also demonstrates superior capability in generating semantically coherent and contextually appropriate oxymoronic pairs. Notably, the same model leads to competitive results also when dealing with difficult and controversial samples (i.e., indirect multimodal oxymorons). Furthermore, the system demonstrates optimal performance (as shown in Table 12)—following closely behind the best Zero-Shot Capt and tied with the tag-based Zero-Shot model—in correctly recognizing the absence of multimodal oxymorons, demonstrating its robustness not only in positive prediction but also in discerning negative cases.

Table 13
Comparison of the different approaches for multimodal oxymoron validation on the overall dataset. **Bold** denotes the best approach for the corresponding subset.

Approach	Input	% Identified Oxymorons	% of Valid Oxymorons
Zero-Shot	Tags	88.19%	38.19%
	Capt	87.64%	33.33%
	Tag+Capt	**89.17%**	39.86%
One-Shot	Tags	88.47%	51.11%
	Capt	87.36%	**58.06%**
	Tag+Capt	87.92%	54.86%

As a result of its favorable combination of performance in recognition and accuracy of the predicted oxymorons, we selected the One-Shot GPT captions-based approach for future in-depth study, providing important insights into the complex landscape of multimodal oxymoron recognition.

5.4 Error Analysis

In order to provide a roadmap for future research directions, a systematic error analysis has been performed to understand if the memes share any common characteristics that need to be properly addressed from now onward. In particular, we analyzed the multimodal oxymoron memes in accordance with the label obtained from the best-performing approach (i.e., GPT_Capt).

5.4.1 Reliance on Prior Knowledge. The analyzed model exhibits a tendency to rely on prior knowledge, resulting in the correct prediction of common or well-known oxymorons, such as *bright-dark*, even when not conveyed in the input. This limits their ability to interpret the content contextually; 17% of the oxymorons predicted on the Italian subset of the dataset, and 9% for the Spanish subset, exhibit this behavior.

For instance, for the first meme reported in Table 14, the model successfully predicted the oxymoron *bright-dark*, matching the majority of annotators' choice *buio-luminoso*. While the model's prediction is accurate, it is essential to highlight a noteworthy aspect of the model's behavior, specifically, the presence of hallucination in its

Table 14
Error analysis.

Hallucination	Hallucination	Source interference
		
Estate agent: As you can see it is a very bright house. The house:	*Estate agent: As you can see it is a very bright house. The house:*	*When you are studying and the sentence 'As we already know' appears in the book.*
Image caption: `a photograph of a living room with a couch and a table.`	Image caption: `a photograph of a lighted house with people walking around it.`	Image caption: `a photograph of a woman with a red nail polish holding a red nail.`
Predicted Oxymoron: `''bright-dark''`	Predicted Oxymoron: `''bright-dark''`	Predicted Oxymoron: `''Alpha-Woman''`

response. The input provided to the model was found to be insufficient for an unequivocal identification of the given oxymoron. In fact, the image caption reported in the prompt did not contain explicit elements that could be directly linked to the concept of *darkness*, even though the image itself represented a dark room. The model's ability to predict the correct oxymoron suggests a potential reliance on prior knowledge, particularly the well-known oxymoron *bright-dark*, which might have influenced its response. This phenomenon raises the question of whether the model's performance is driven by a learned association rather than a direct interpretation of the provided input.

A parallel observation is noted in the second meme in Table 14 with identical text but a different image portraying a luminous house. In this case, no oxymoron should be identified as both the text and image convey the same concept of *brightness*. However, the model predicts the oxymoron *bright-dark*. This recurrence of the same oxymoron, despite the absence of contrasting elements in the input, indicates a consistent reliance on pre-existing associations. This highlights a concerning trend in the model's behavior, where it tends to inject familiar oxymoronic pairs even when not contextually appropriate, only because one of the concepts that compose it appears in the input.

5.4.2 Influence of Source Information. Another noteworthy phenomenon is that, in certain instances, the meme's source (expressed via a logo or link) influences the oxymoron detection process. While including the meme's source (e.g., *alpha woman*) in the text description can be valuable for downstream analyses—such as identifying sources that frequently share oxymoronic memes—it poses a challenge during the prediction phase.

As previously mentioned in Section 4.1 regarding dataset composition, memes often contain embedded textual elements. Specifically, 28% of the dataset includes text related to the meme's source or its creation/sharing. These elements are considered by the models—either by vision-language models treating them as part of the image or by LLMs interpreting them as part of the caption or transcription—which can lead to misclassifications or the generation of incorrect labels. Although, in most cases, ChatGPT is able to discern and exclude source-related information during oxymoron detection, there are instances in which such information leads to model errors.

A representative example can be found in the third meme in Table 14, where the model erroneously predicts alpha-woman as an oxymoron.[13] In such cases, although the text contains useful metadata regarding the meme's origin, the model fails to filter out this content during oxymoron detection. As a result, it may generate what can be described as a multimodal oxymoron, in which one of the concepts originates from the meme's source rather than the intended semantic content.

5.4.3 Incomplete or Incorrect Captions. Another salient phenomenon we identified is the occurrence of incomplete or incorrectly generated captions. In particular, incomplete captions refer to those that, while accurately describing the visual content of the scene, fail to include key semantic elements that are crucial for understanding the opposition underlying the oxymoron. For instance, in the caption "a photograph of a woman in a dress sitting next to a vase of flowers", the description is visually correct but omits the emotional state (sadness) which is essential for recognizing the multimodal oxymoron

13 We acknowledge the potential misogynistic connotations of the identified oxymoron. However, our aim is to highlight a specific aspect of the model's behavior. The recognition and mitigation of potentially harmful or stereotypical outputs in language models is an important research area, which we defer to domain experts.

Table 15
Error analysis.

Incomplete Caption	Missing Reasoning
Usciamo a prendere un po' d'aria fresca 	Io: sono competitivo ma gioco onestamente. Sempre io che compro Parco della vittoria a mio cugino 
Let's go out and get some fresh air	*Me: I am competitive but I play honestly* *Always me buying my cousin Victory Park*

Image caption:
`a photograph of a car that has been damaged and is parked.`
Predicted Oxymoron:
`''outing - damage''`
	Image caption:
	`a photograph of a man shaking another man's hand with money.`
	Predicted Oxymoron:
	`''competitiveness - honesty''`

based on the contrast "happy–sad". These cases constitute the majority of captioning issues observed. Incorrect captions, on the other hand, involve misrepresentations of the visual content or the inclusion of details not present in the image. These are relatively rare, accounting for only about 5% of the oxymorons in the dataset. An example is "a photograph of a man doing a kick with a woman", which incorrectly describes an image where a man is actually lifting or holding a plus-sized woman, not performing a kick. In both cases, the prompt generated by the inclusion of such captions might result in an unreliable oxymoron identification. Considering the Italian and the Spanish subsets of data, for 55% and 45% of oxymoron memes, respectively, the captions are incomplete or erroneous, and consequently, the model performs a wrong prediction. For example, in the first meme reported in Table 15, within the image caption, a car is denoted as *damaged*. However, this description fails to establish any discernible relationship with the concept of *hot*, making it impractical for the model to accurately predict the corresponding oxymoron (*cold-hot*). This underscores the importance of addressing challenges related to caption accuracy, especially when depicting abstract concepts, as inaccuracies can significantly impede the model's ability to comprehend and identify multimodal oxymorons.

5.4.4 Missing Reasoning. Another noteworthy phenomenon is observed in certain instances where, despite the model having sufficient information to accurately detect the "correct" oxymoron, it predicts an incorrect oxymoron influenced by something explicitly mentioned in the text.

In 5% of oxymoron memes in the Italian subset of data, and in the 9% of oxymoron memes in the Spanish subset, although the prompt includes all the information necessary for the correct prediction, a wrong concept is identified as part of the oxymoron. For instance, in the second meme reported in Table 15, the model predicts *competitiveness - honesty*, which is not an oxymoron, seemingly driven by the presence of both terms in

the text. This occurs despite the explicit instruction in the prompt to look for multimodal oxymorons, where the concepts are distributed in the two modalities. The "correct" oxymoron, *honesty-corruption*, is deducible from the information in the input since *onesto* "honest" is explicitly mentioned in the text, and the image caption references *shaking hand with money*, a well-known symbol of corruption. This behavior suggests the need to strengthen the relationship with the prompt in order to ensure that all expressed conditions are met.

5.4.5 Difficulty in Discerning Contrastive Concepts. A final notable aspect to highlight is the model's inability to discern contrastive concepts that could form an oxymoron. From the performed analysis emerges that models face difficulties in discerning antonymic relationships between contrasting concepts, hindering their capability to identify oxymorons accurately. This is evident in cases where the model incorrectly associates unrelated terms, resulting in inaccurate predictions.

We see that 28% of the oxymorons predicted on the Italian subset of the dataset, and 29% of the Spanish subset, are associated with a couple of concepts that don't represent an oxymoron due to an error of the model in discerning contrastive concepts. For instance, in the aforementioned example (first meme of Table 15), the model incorrectly perceives the concept of *damaged* as opposing the concept of *outing*, leading to the erroneous prediction of the oxymoron *outing - damage*. This reveals a broader limitation in the model's ability to accurately identify what constitutes an antonymic relationship, underscoring the need for improved representation of semantic information—especially the semantic-lexical relationships essential for the precise handling of both textual and multimodal oxymorons.

To answer **RQ3**, while some models can successfully predict multimodal oxymorons, several challenges and limitations persist. Challenges include reliance on prior knowledge, susceptibility to source-related influences, inaccuracies in captions, difficulty discerning contrastive concepts, and the implicit expression of knowledge, leading to potential mispredictions. Further research is needed to address these limitations and enhance models' accuracy in multimodal oxymoron related tasks.

6. Meme Generation

A pipeline for automatically generating multimodal oxymorons is proposed. A key element of the proposed approach is the inclusion of generative large language models. The generation of all the parts that compose a meme ensures that the final meme covers all the characteristics that define a multimodal oxymoron. The proposed pipeline is schematized in Figure 5 and composed of the following steps:

- **Oxymoron selection**: The first step of the proposed approach focuses on the selection of an existing oxymoron from a list of validated ones (Bolognesi et al. 2024; La Pietra and Masini 2020).[14] The selected oxymoron will be used for the generation of a meme that contains it in a

14 The selected oxymorons include the following: *dishonesty - sincerity, formality - informality, lightness - heaviness, order - disorder, abundance - misery, active - inactive, boring - stimulating, immature - mature, generosity - greed, ugliness - beauty, hot - cold, sincere - liar, dirty - clean, inefficient - efficient, happiness - unhappiness, light - dark, immature - wise, young - old, strength - weakness, simple - difficult, able - unable, shortness - height, sharpness - dullness, rich - poor, comfortable - uncomfortable.*

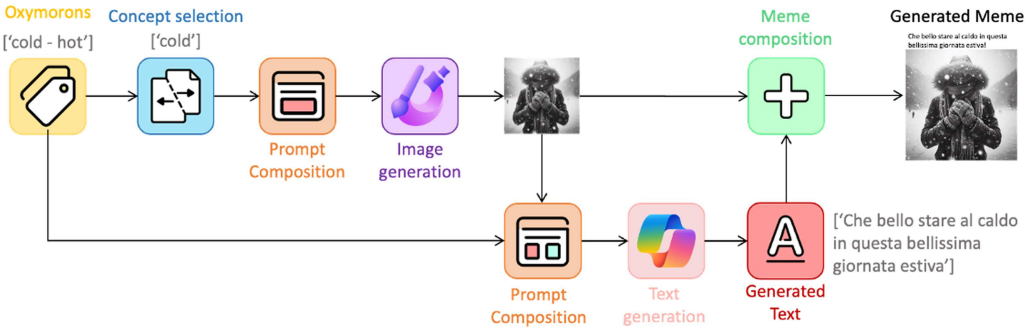


Figure 5
Schematic representation of the proposed pipeline for multimodal oxymoron meme generation.

multimodal form. Note that, while the selected oxymorons will be adopted for the generation of the memes, different perceptions can lead to different oxymorons. However, the proposed pipeline aims to define a multimodal oxymoron in which the selected couple of concepts, in accordance with the main interpretation of the meme, compose a multimodal oxymoron.

- **Image generation**: As previously described, a given oxymoron can be represented in multimodal content in different ways. To ensure the multimodal representation of the oxymoron itself, one of the concepts that compose it has been selected for the generation of the image. In particular, given the two concepts that compose the oxymoron, the first one has been isolated and used in image generation. The prompt for the image generation is composed as follows: *"a situation commonly associated with the concept of [concept]"*. For the generation process, we adopted Copilot Image Creator.[15] Copilot Image Creator is a model based on DALL-E 3[16] able to generate images from text expressed in more than 100 languages.
- **Text generation**: An additional prompt to generate a text to compose a meme with the characteristics necessary to express the concepts in a multimodal oxymoron has been defined. Several prompts of different types and with different inputs and instructions were investigated. Additional information and examples of the prompts considered are reported in Appendix 10. A qualitative analysis on a subset of well known oxymorons was performed to identify the most appropriate prompt for multimodal oxymoron generation. The selected prompt not only includes the concept that composes the oxymoron that we want to represent, but also the previously generated image. The inclusion of the image in the prompt definition allows the model to refer, within the generated text, to the elements that compose it, resulting in a more realistic meme. The prompt adopted for text generation is reported in

15 Service available at `https://copilot.microsoft.com/images/create?FORM=GENEXP`. Accessed between January and March 2024.
16 `https://openai.com/dall-e-3`.

Appendix 10, while Table 16 reports its English translation, highlighting the parts that compose it.

- **Meme composition**: The generated elements (image and text) have been finally merged into a single representation to create the final meme.

The proposed procedure allows the generation of multimodal oxymorons. While the proposed methodology has been defined, starting from the definition of a multimodal oxymoron, additional constraints have been included to improve the quality of the final result. In particular, the first and fourth constraints invite the model in the generation of a direct multimodal oxymoron. Those constraints have been included to evaluate the ability of the model in the generation of the most common and easy type of multimodal oxymorons, also considering the limitations of large language models that emerged from the analysis reported in Section 5.4. Additionally, the fourth constraint has been introduced to limit the tendency of the generative model to report both concepts in the text.

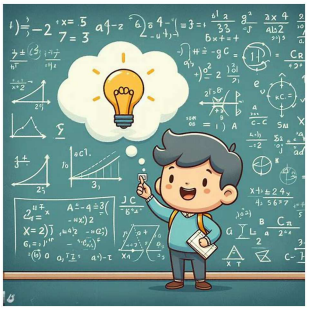
6.1 Synthetic Multimodal Oxymorons Dataset

The proposed pipeline has been adopted in the generation of 30 multimodal oxymorons in Italian. The generated multimodal oxymorons have been validated by domain

Table 16
Prompt for text generation.

	Text Generation Prompt
Instruction	Generate a sentence that, in association with the provided image, represents a meme whose irony is based on the juxtaposition of contrasting concepts.
Oxymorons Characteristics	Given CI, the concept represented by the provided image, and CT, the opposite concept for text generation, generate a text for a meme that explicitly contains CT and meets all of the following characteristics: (i) The text must express CT explicitly. (ii) The text must express CT indicated in a way that coexists in time with the concept represented by the image (CI) (i.e., it must not contain terms such as "before" or "after"). (iii) The text must express CT indicated in reference to the same subject as the image. (iv) The text must not contain other abstract concepts: it must not state either CI or other concepts opposite to CT.
Expected Output	Answer only with the generated text.
Example	For example, given an image of a woman crying, representing the concept of "sadness," the text, in relation to the opposite concept of "happiness", could be "The happiness that sets me apart" or "The happiest day of my life".
Input Definition	Generate the meme text based on the following concepts: CI = [*concept-1*] CT = [*concept-2*]

Table 17
Examples of generated memes. The English translation of the original text is also reported.

Generated memes	
La mia sincerità è la mia forza.	La complessità delle mie idee
	
My sincerity is my strength.	*The complexity of my ideas.*

experts; all the generated multimodal oxymorons are considered valid in accordance with the definition provided in Section 3. Examples of the generated multimodal oxymorons are reported in Table 17.

To evaluate if the generated memes were perceived as multimodal oxymorons, we conducted a two-step annotation process using crowdsourced participants (native speakers), collecting five responses per oxymoron.[17] In the first step, participants were asked to indicate the general concept they associated with the image. Additionally, similar to the procedure used in the dataset validation, they were invited to label the multimodal oxymoron, responding to the following open-ended prompt: "If you believe this meme contains one (or more) multimodal oxymorons, please indicate the pair of contradictory concepts it expresses. You may leave this field blank if none are present". In the second step, participants were asked to rate, on a 7-point Likert scale:

- How well the image represented the target concept used in the meme's generation;
- How well the meme as a whole represented the target opposition.

Memes were considered to be clearly perceived as multimodal oxymorons if they scored above the midpoint of the scale for both image representation and overall meme effectiveness in conveying the target oxymoron. The results indicate that, while the generated oxymorons conform to the expert definition, where the text conveys one concept and the image conveys its opposite, both referring to the same entity simultaneously, only 83% of the generated memes are clearly recognized as multimodal oxymorons by annotators. Only 3 out of 30 images were not considered good representations of the target concept.

17 No distractors were included in the annotation process, which could be considered a limitation of the results.

For example, the first meme reported in Table 17 is clearly perceived as a multimodal representation of the oxymoron "*sincerità - disonestà*" (*"sincerity - dishonesty"*), while the second meme might be harder to detect due to the image perception. The image that composes the multimodal oxymoron, despite being created for the concept of "*simple*", is more frequently associated with different and contradictory concepts (i.e., "intelligence", "idea", "logic", "genius"). Therefore, while the meme reflects all the characteristics necessary to be considered a multimodal oxymoron, the most common perception leads to an interpretation of the conveyed message as a non-oxymoron. In fact, only the (uncommon) perception of the image as "*simple*" or "*easy*" leads to the identification of contradictory concepts and, therefore, to the multimodal oxymoron identification.

In relation to **RQ4**, we can say that the proposed pipeline represents an effective automatic approach for generating a valid multimodal representation of existing oxymorons. However, as an additional improvement of the pipeline, to guarantee the generation of more realistic and clear multimodal oxymoron memes (i.e., in which the concepts that represent the oxymoron correspond with the most common perception), we suggest the inclusion of a validation process of the generated images to assess their efficacy in conveying the related concept.

7. Conclusions and Future Work

This article introduces a novel phenomenon that has not yet been systematically explored: the multimodal oxymoron. Like their textual counterparts, multimodal oxymorons have been largely overlooked in both linguistic and computational research. We define their characteristics within the current state of the art, distinguishing them not only from textual and visual oxymorons but also from related figures of speech such as antithesis and paradox.

A central contribution of this work is the Multimodal Oxymorons dataset (MOxy), consisting of 300 memes in Italian and Spanish. To our knowledge, this represents the first resource specifically dedicated to multimodal oxymorons, paving the way for research on their perception and for evaluating the ability of machine learning models to identify them. Beyond the dataset itself, we propose a framework that facilitates expansion with additional samples, can be adapted to new languages, and includes a pipeline for the automatic generation of multimodal oxymorons.

Our preliminary experiments with automatic prediction highlight several challenges for current vision-language and large language models, across different evaluation setups. These include misinterpretation of antonyms, reliance on spurious visual cues, and modality-specific errors, reflecting broader limitations in reasoning across modalities.

Future research will extend our work in several directions. First, we plan to broaden the linguistic scope by creating resources in English and to investigate the impact of using language-specific prompts and captioning models, especially when dealing with specific linguistic and multimodal phenomena such as oxymorons. Second, we aim to evaluate a wider range of models, both open-source and proprietary, spanning diverse parameter scales and language coverage, to better assess generalizability. This will also involve probing the reasoning capabilities of multiple multimodal models (e.g., GPT-4o-mini) in a multimodal space. Third, we will expand experiments with automatically generated multimodal oxymorons and continue refining the proposed pipeline.

Finally, building on insights from the literature, we recognize that memes are often used to convey hateful content toward specific targets (Kiela et al. 2020; Fersini et al. 2022), yet existing approaches frequently fail to capture multimodal figures of speech such as metaphors (Rizzi et al. 2023). Future work will therefore extend the study of multimodal figurative expressions—including both metaphors and oxymorons—with the goal of improving the ability of classification models to handle these complex phenomena.

8. Limitations

We recognize that the size of the dataset can be seen as a limitation: The manually curated subset consists of 300 instances, while the automatically generated multimodal subset includes only 30 oxymorons. This reflects a deliberate choice to prioritize quality over quantity, with the primary goal of providing a replicable methodology for systematically constructing oxymorons and investigating an underexplored linguistic phenomenon, and it is consistent with previous work on antonyms and oxymorons. Nonetheless, the limited size remains a constraint. In future work, we plan to address this by extending the dataset, evaluating a larger number of examples, and ultimately moving toward the creation of a large-scale training resource.

Our experiments are conducted with a single proprietary LLM (GPT-4o-mini). This choice was motivated by computational constraints, but we acknowledge that including a broader mix of open-source and proprietary models with varying parameter sizes and language coverage would yield a more comprehensive evaluation. At the same time, we believe that the proposed methods should be robust across model families, sizes, and languages.

We highlight both limitations as concrete directions for future work.

9. Multimodal Oxymoron Identification

To explore the complex interpretation of multimodal oxymorons, several labels to represent different perspectives have been collected via a crowdsourcing platform. Considering the novelty aspects represented by multimodal oxymorons themselves and the peculiar characteristics that distinguish them, an introductory part has been included to provide guidelines and definitions. To better assess the presence of such figurative language device, native speakers have been involved, and the labeling form has been translated (and validated) consequently. In particular, in the first part of the form, a general introduction to the problem is given to make the annotators aware of the ongoing research and the potential exposure to harmful content. After gathering the annotator's consent in the participation, the guidelines reported in Table 18 are provided (for the sake of completeness and accessibility, the English translation of the guidelines is also reported in Table 19).

Finally, as shown in Figure 6, the actual labeling phase is composed of the meme to be evaluated encapsulated in the following questions:

- Secondo te, questo meme rappresenta un ossimoro multimodale? [*In your opinion, does this meme represent a multimodal oxymoron?*]
- Se si, a quale coppia di opposti associ il meme? [*If so, which pair of opposing concepts do you associate the meme with?*]

Table 18
Original annotation guidelines for Italian annotators.

	Annotation guidelines
Definition	Per ossimoro multimodale intendiamo in questo caso l'accostamento di elementi contrari o contraddittori in una sola rappresentazione, ad esempio "felice infelicità", "amara dolcezza", "follia lucida".
Instruction	Guarda l'immagine, poi in base alla definizione di ossimoro fornita rispondi alla domanda. Se nell'immagine riscontri la presenza di un ossimoro indica nella seconda domanda i concetti opposti che rivedi nell'immagine.
Example	Ad esempio: 1) Viene fornito un meme il cui testo recita *"Sono felice di andare all'appuntamento" e l'immagine mostra il volto di una donna che piange.* Se pensi che l'immagine rappresenti un ossimoro, rispondi "Sì" alla domanda "Secondo voi, questo meme rappresenta un ossimoro?". Si noti che **il concetto può essere chiaramente affermato o implicito.** Assicurati però che i concetti opposti non si riferiscano a eventi/tempi diversi; **i concetti opposti devono coesistere nel meme.** *Nella seconda domanda ti viene chiesto di associare la coppia di concetti contrastanti che hai visto nell'immagine al meme.*
Expected Output	I concetti possono essere espressi in modi diversi: ✓felicità e tristezza ✓felicità, tristezza ✓felicità triste ✓felice triste ✓felicità triste ✓felice triste ✗faccia felice - triste ✗faccia sorridente - faccia triste Non è necessario descrivere l'immagine, ma indicare i concetti contrastanti. Se pensi che un meme rappresenti più ossimori, elencali separandoli con una virgola.
Example	2) Ti viene dato un meme il cui testo dice "l'estate più fredda degli ultimi 10 anni" e l'immagine mostra un paesaggio innevato.
Expected Output	In questo caso non ci sono due concetti opposti, quindi si può rispondere "No" alla prima domanda e lasciare in bianco la seconda. Se avete risposto "No", potete passare al meme successivo.

10. Multimodal Oxymoron Generation

The prompt adopted for generating the text to include within the meme in order to create a multimodal oxymoron is the following:[18]

[ITA] *Genera una frase che, in associazione all'immagine fornita, rappresenta un meme la cui ironia si basa sulla giustapposizione di concetti contrastanti.*

18 We report the original prompt followed by its English translation.

Table 19
Annotation guidelines (English translation).

	Annotation guidelines
Definition	By multimodal oxymoron, we refer here to the juxtaposition of contradictory or opposite elements within a single representation, such as "happy sadness," "bitter sweetness," or "lucid madness."
Instruction	Look at the image, then based on the provided definition of an oxymoron, answer the question. If you find an oxymoron in the image, indicate in the second question the opposing concepts you see in the image.
Example	For instance: 1) A meme is provided where the text reads *"I am happy to go to the appointment" and the image shows the face of a crying woman.* If you think the image represents an oxymoron, answer "Yes" to the question "Do you think this meme represents an oxymoron?". Note that **the concept may be clearly stated or implied.** However, make sure the opposing concepts do not refer to different events/times; **the opposing concepts must coexist in the meme.** *In the second question, you are asked to associate the pair of contrasting concepts you observed in the image with the meme.*
Expected Output	The concepts can be expressed in different ways: ✓happiness and sadness ✓happy, sad ✓happy sadness ✓happy sad ✓sad happiness ✓sad happy ✗happy face - sad face ✗smiling face - sad face It is not necessary to describe the image, but to indicate the contrasting concepts. If you think a meme represents multiple oxymorons, list them separated by a comma.
Example	2) You are given a meme where the text says "the coldest summer in the last 10 years" and the image shows a snowy landscape.
Expected Output	In this case, there are no two opposing concepts, so you can answer "No" to the first question and leave the second blank. If you answered "No," you can move on to the next meme.

Dati CI, concetto rappresentato dall'immagine fornita, e CT, concetto opposto per la generazione del testo, genera un testo per un meme che contenga esplicitamente CT in che rispetti tutte le seguenti caratteristiche:

- *Il testo deve esprimere CT in modo esplicito.*
- *Il testo deve esprimere CT indicato in modo che coesista nel tempo con il concetto rappresentato dall'immagine (CI) (i.e., non deve contenere termini come "prima" o "dopo").*
- *Il testo deve esprime CT indicato in riferimento allo stesso soggetto dell'immagine.*

Figure 6
An extract of the labeling form.

- *Il testo non deve contentere altri concetti astratti: non deve riportare nè CI nè altri concetti opposti a CT.*

Rispondi solo con il testo generato.
Ad esempio, data l'immagine di una donna che piange, rappresentante il concetto di "tristezza", il testo, in relazione al concetto opposto di "felicità", potrebbe essere "la felicità che mi contraddistingue" o "Il giorno più felice della mia vita".
Genera il testo del meme basandoti sui seguenti concetti:
CI = "bello",
CT = "mostruoso".

[ENG] *Generates a sentence that, in association with the provided image, represents a meme whose irony is based on the juxtaposition of contrasting concepts. Given CI, the concept represented by the provided image, and CT, the opposing concept for text generation, generates a text for a meme that explicitly contains CT in which all of the following characteristics are met:*

- *The text must express CT explicitly. text must express CT indicated in such a way that it coexists in time with the concept represented by the image (CI) (i.e., it must not contain terms such as 'before' or 'after'). text must express CT indicated with reference to the same subject as the image. text must not contain*

other abstract concepts: it must not contain either CI or other concepts opposite CT.

Reply only with the generated text. For example, given an image of a woman crying, representing the concept of 'sadness', the text, in relation to the opposite concept of 'happiness', could be 'the happiness that characterises me' or 'the happiest day of my life'. Generate the text of the meme based on the following concepts: CI = 'beautiful', CT = 'monstrous'.

10.1 Alternative Prompts Considered in the Evaluation Phase

Several prompts have been defined and evaluated for the generation of multimodal oxymorons. Examples of alternative prompts are summarized in Tables 20 and 21. The first prompt (Table 20), different from the other ones that have been proposed, doesn't include the complete oxymoron, but only reports the concept adopted for image generation. This prompt consequently leaves to the model the choice of the concept in contrast with the image in the generation of the oxymoron. This freedom does not guarantee, however, that the generated oxymoron corresponds to the one selected in the generation pipeline.

The second prompt (Table 21), instead, provides a more guided and controlled instruction to the model by explicitly reporting the contrastive contents that should compose the generated oxymoron. Different from the previous prompt, in this case, the generated image is also provided as part of the input along with the concept employed in the generation phase.

11. Hyperparameter Configuration

We report the adopted hyperparameters configuration along with implementation details to allow for the reproducibility of the results.

Table 20
Prompt based on image-concept only. The model is requires to select the textual concept to generate the oxymoron.

	Text Generation Prompt
Instruction	Generate a sentence that, in association with the image provided, represents a meme whose irony is based on the juxtaposition of contrasting concepts.
Example	For example, given the image of a woman crying, the text could be "the happiness that distinguishes me", based on the contrasting concepts of "happiness" and "sadness".
Expected Output	Generate a funny text based on a concept opposite to that represented by the image.
Oxymoron Characteristics	Express it, in reference to the same subject of the image, so that the two concepts coexist over time. The text combined with the image must create irony.
Input Definition	The image represents the concept of [*concept*]

Table 21
Prompt explicitly reporting the concepts that should compose the oxymoron.

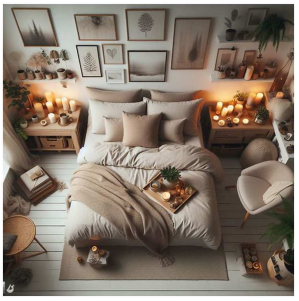
	Text Generation Prompt
Instruction	Generate a sentence that, in association with the image provided, represents a meme whose irony is based on the juxtaposition of contrasting concepts.
Example	For example, given the image of a woman crying, the text could be "the happiness that distinguishes me", based on the contrasting concepts of "happiness" and "sadness".
Input Definition	The attached image represents the concept of "[*concept-1*]". Generate a text based on the contrasting concept of "[*concept-2*]". Oxymorons.
Characteristics	Express it so that the two concepts coexist over time in reference to the same subject of the image. The text combined with the image must create irony.
Expected Output	Respond only with the generated text.

For what concerns the vision-language models, we fine-tuned `Gregor/mblip-mt0-xl` and `sentence-transformers/CLIP-ViT-B-32-multilingual-v1` as available on HuggingFace. Both models were fine-tuned adopting default parameters as suggested by the authors, with a 10-fold cross validation approach on the MOxy dataset for 5 epochs and a batch size of 4.

12. Generated Memes

Further examples of generated memes are reported in Table 22.

Table 22
Further examples of generated memes along with the antonyms adopted for the generation.

Generated memes		
La ricchezza che mi circonda. 	"Il fastidio di dormire in questo letto." 	La bellezza che mi definisce. 
The wealth that surrounds me.	*"The discomfort of sleeping in this bed."*	*The beauty that defines me.*
Original Oxymoron:		
‘‘rich-poor’’	‘‘comfortable -uncomfortable’’	‘‘beauty-ugliness’’
Concepts adopted for image generation:		
‘‘poor’’	‘‘comfortable’’	‘‘ugliness’’

Generated memes pt.2		
La generosità nelle mie mani. 	L’entusiasmo che mi anima ogni giorno. 	La mia bassa statura è inconfondibile. 
Generosity in my hands.	*The enthusiasm that animates me every day.*	*My short stature is unmistakable.*
Original Oxymoron:		
‘‘avarice - generosity’’	‘‘enthusiasm - boredom’’	‘‘tall - short’’
Concepts adopted for image generation:		
‘‘avarice’’	‘‘boredom’’	‘‘tall’’

Acknowledgments

We acknowledge the support of the PNRR ICSC National Research Centre for High Performance Computing, Big Data and Quantum Computing (CN00000013), under the NRRP MUR program funded by the NextGenerationEU. The work of Elisabetta Fersini has been partially funded by MUR under the grant ReGAInS, Dipartimenti di Eccellenza 2023-2027 of the Department of Informatics, Systems and Communication at the University of Milano-Bicocca. The work of Paolo Rosso was in the framework of the FairTransNLP-Stereotypes research project, grant PID2021-124361OB-C31 funded by MCIN/AEI/10.13039/501100011033 and by ERDF/EU.

References

Athanasiadou, Angeliki. 2024. On the margins of figurative thought and language. *Lingua*, 299:103655. https://doi.org/10.1016/j.lingua.2023.103655

Barthes, R. and S. Heath. 1977. *Image, Music, Text*. Fontana Press.

Benammar, Silya. 2023. Multimodal figuration in internet memes. *metaphorik.de*, 34.

Bolognesi, Marianna M., Claudia Roberta Combei, Marta La Pietra, and Francesca Masini. 2024. What makes an awfully good oxymoron? *Language and Cognition*, 16(1):242–262. https://doi.org/10.1017/langcog.2023.68

Cerini, Ludovica, Eliana Di Palma, and Alessandro Lenci. 2022. From speed to car and back: An exploratory study about associations between abstract nouns and images, Dobnik, Simon. In *Proceedings of the 2022 CLASP Conference on (Dis)embodiment*, pages 80–88.

Choi, Wan Chong and Chi In Chang. 2025. A survey of techniques, key components, strategies, challenges, and student perspectives on prompt engineering for large language models (LLMs) in education. https://doi.org/10.20944/preprints202503.1808.v1

Dancygier, Barbara and Eve Sweetser. 2014. *Figurative Language*, Cambridge University Press.

De Mauro, Tullio. 2016. Il nuovo vocabolario di base della lingua italiana. *Internazionale*. https://www.internazionale.it/opinione/tullio-de-mauro/2016/12/23/il-nuovo-vocabolario-di-base-della-lingua-italiana

Dynel, Marta. 2016. Two layers of overt untruthfulness: When irony meets metaphor, hyperbole or meiosis. *Pragmatics & Cognition*, 23(2):259–283. https://doi.org/10.1075/pc.23.2.03dyn

Fahnestock, Jeanne. 1999. In *Rhetorical Figures in Science*. Oxford University Press. https://doi.org/10.1093/oso/9780195117509.001.0001

Fersini, Elisabetta, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 Task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549. https://doi.org/10.18653/v1/2022.semeval-1.74

Flayih, Reja'a M. 2009. A linguistic study of oxymoron. *Journal of Kerbala University*, 7(3):30–40.

Forceville, Charles. 2006. Non-verbal and Multimodal Metaphor in a Cognitivist Framework: Agendas for Research. De Gruyter Mouton. https://doi.org/10.1515/9783110197761.5.379

Forceville, Charles and Eduardo Urios-Aparisi. 2009. *Multimodal Metaphor*, volume 11, Walter de Gruyter. https://doi.org/10.1515/9783110215366

Geigle, Gregor, Abhay Jain, Radu Timofte, and Goran Glavaš. 2024. mBLIP: Efficient bootstrapping of multilingual vision-LLMs. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 7–25. https://doi.org/10.18653/v1/2024.alvr-1.2

Gibbs, R. W. Jr. and L. R. Kearney. 1994. When parting is such sweet sorrow: The comprehension and appreciation of oxymora. *Journal of Psycholinguistic Research*, 23(1):75–89. https://doi.org/10.1007/BF02143177

Jahameh, Haifaa and Aseel Zibin. 2023. The use of monomodal and multimodal metaphors in advertising Jordanian and American food products on Facebook: A comparative study. *Heliyon*, 9(5). https://doi.org/10.1016/j.heliyon.2023.e15178, PubMed: 37131431

Jones, S. 2003. *Antonymy: A Corpus-Based Perspective*. Routledge Advances in Corpus Linguistics. Taylor & Francis. https://doi.org/10.4324/9780203166253

Kiela, Douwe, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide

Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems*, 33:2611–2624.

Kress, Gunther and Theo Van Leeuwen. 2020. *Reading Images: The Grammar of Visual Design*. Routledge. https://doi.org/10.4324/9781003099857

La Pietra, Marta and Francesca Masini. 2020. Oxymorons: A preliminary corpus investigation. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 176–185. https://doi.org/10.18653/v1/2020.figlang-1.24

Lakoff, G. and M. Johnson. 1980. *Metaphors We Live By*. University of Chicago Press.

Li, Junnan, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*. https://doi.org/10.48550/ARXIV.2201.12086

Lou, Adrian. 2017. Multimodal simile: The "when" meme in social media discourse. *English Text Construction*, 10(1):106–131. https://doi.org/10.1075/etc.10.1.06lou

Nadeau, Claude and Yoshua Bengio. 1999. Inference for the generalization error. *Advances in Neural Information Processing Systems*, 12.

Nadeau, Claude and Yoshua Bengio. 2003. Inference for the generalization error. *Machine Learning*, 52(3):239–281. https://doi.org/10.1023/a:1024068626366

Radford, A., J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763.

Reimers, Nils and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/D19-1410

Richard, I. A. 1936. *The Philosophy of Rhetoric*. Oxford University Press.

Rizzi, Giulia, Francesca Gasparini, Aurora Saibene, Paolo Rosso, and Elisabetta Fersini. 2023. Recognizing misogynous memes: Biased models and tricky archetypes. *Information Processing & Management*, 60(5):103474. https://doi.org/10.1016/j.ipm.2023.103474

Ruiz, Javier Herrero. 2009. *Understanding Tropes: At the Crossroads between Pragmatics and Cognition*, volume 75. Peter Lang.

Ruiz de Mendoza Ibáñez, Francisco José. 2020. Understanding figures of speech: Dependency relations and organizational patterns. *Language & Communication*, 71:16–38. https://doi.org/10.1016/j.langcom.2019.12.002

Ruiz de Mendoza Ibáñez, Francisco José and Alicia Galera-Masegosa. 2012. Modelos cognitivos, operaciones cognitivas y usos figurados del lenguaje. *Forma y Función*, 25(2):11–38.

Schulhoff, Sander, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, H. Han, Sevien Schulhoff, and others. 2024. The prompt report: A systematic survey of prompting techniques. *arXiv preprint arXiv: 2406.06608*, 5.

Scott, Kate. 2021. Memes as multimodal metaphors: A relevance theory analysis. *Pragmatics & Cognition*, 28(2):277–298. https://doi.org/10.1075/pc.21010.sco

Shen, Yeshayahu. 1987. On the structure and understanding of poetic oxymoron. *Poetics Today*, 8(1):105–122. https://doi.org/10.2307/1773004

Tomás, David, Reynier Ortega-Bueno, Guobiao Zhang, Paolo Rosso, and Rossano Schifanella. 2022. Transformer-based models for multimodal irony detection. *Journal of Ambient Intelligence and Humanized Computing*, 14(6):7399–7410. https://doi.org/10.1007/s12652-022-04447-y

Tseronis, Assimakis and Charles Forceville. 2017. The argumentative relevance of visual and multimodal antithesis in Frederick Wiseman's documentaries. In Assimakis Tseronis and Charles Forceville, editors, *Multimodal Argumentation and Rhetoric in Media Genres*, Johns Benjamins, pages 165–188. https://doi.org/10.1075/aic.14.07tse

Turpin, Miles, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965. https://doi.org/10.52202/075280-3275

Veale, Tony. 2008. Figure-ground duality in humour: A multi-modal perspective. *Lodz*

Papers in Pragmatics, 4(1). https://doi.org/10.2478/v10016-008-0009-z

Veale, Tony, Ekaterina Shutova, and Beata Beigman Klebanov. 2016. Metaphor: A computational perspective. *Synthesis Lectures on Human Language Technologies*, 9(1):1–160. https://doi.org/10.1007/978-3-031-02160-2

Yosef, Ron, Yonatan Bitton, and Dafna Shahaf. 2023. IRFL: Image recognition of figurative language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1044–1058. https://doi.org/10.18653/v1/2023.findings-emnlp.74

Youden, William J. 1950. Index for rating diagnostic tests. *Cancer*, 3(1):32–35. https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3, PubMed: 15405679

Yus, F. 2023. *Pragmatics of Internet Humour*. Springer International Publishing. https://doi.org/10.1007/978-3-031-31902-0

Zhang, Dongyu, Minghao Zhang, Heting Zhang, Liang Yang, and Hongfei Lin. 2021. MultiMET: A multimodal dataset for metaphor understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3214–3225. https://doi.org/10.18653/v1/2021.acl-long.249

Zhang, Youcai, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. 2023. Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514*. https://doi.org/10.1109/CVPRW63382.2024.00179

Zheng, Tianshi, Yixiang Chen, Chengxi Li, Chunyang Li, Qing Zong, Haochen Shi, Baixuan Xu, Yangqiu Song, Ginny Y. Wong, and Simon See. 2025. The curse of CoT: On the limitations of chain-of-thought in in-context learning. *arXiv preprint arXiv:2504.05081*.

Zhong, Zenan, Suijun Wen, and Shukun Chen. 2023. Research trends in multimodal metaphor: A bibliometric analysis. *Frontiers in Psychology*, 14. https://doi.org/10.3389/fpsyg.2023.1144725, PubMed: 37138971