IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 14 July 2024, accepted 9 September 2024, date of publication 24 September 2024, date of current version 8 November 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3467062

SURVEY

Multimodal Explainable Artificial Intelligence: A Comprehensive Review of Methodological Advances and Future Research Directions

NIKOLAOS RODIS[1], CHRISTOS SARDIANOS[1], PANAGIOTIS RADOGLOU-GRAMMATIKIS[2,3], (Member, IEEE), PANAGIOTIS SARIGIANNIDIS[2], (Member, IEEE), IRAKLIS VARLAMIS[1], (Member, IEEE), AND GEORGIOS TH. PAPADOPOULOS[1], (Member, IEEE)

[1]Department of Informatics and Telematics, Harokopio University of Athens, Attica, 176 76 Athens, Greece
[2]Department of Electrical and Computer Engineering, University of Western Macedonia, 501 50 Kozani, Greece
[3]K3Y, Studentski, Sofia, 1700 Sofia City Province, Bulgaria

Corresponding author: Georgios Th. Papadopoulos (g.th.papadopoulos@hua.gr)

This work was supported by European Union's Horizon Europe Research and Development Program (ONELAB) under Grant 101073924.

ABSTRACT Despite the fact that Artificial Intelligence (AI) has boosted the achievement of remarkable results across numerous data analysis tasks, however, this is typically accompanied by a significant shortcoming in the exhibited transparency and trustworthiness of the developed systems. In order to address the latter challenge, the so-called eXplainable AI (XAI) research field has emerged, which aims, among others, at estimating meaningful explanations regarding the employed model's reasoning process. The current study focuses on systematically analyzing the recent advances in the area of Multimodal XAI (MXAI), which comprises methods that involve multiple modalities in the primary prediction and explanation tasks. In particular, the relevant AI-boosted prediction tasks and publicly available datasets used for learning/evaluating explanations in multimodal scenarios are initially described. Subsequently, a systematic and comprehensive analysis of the MXAI methods of the literature is provided, taking into account the following key criteria: a) The number of the involved modalities (in the employed AI module), b) The processing stage at which explanations are generated, and c) The type of the adopted methodology (i.e. the actual mechanism and mathematical formalization) for producing explanations. Then, a thorough analysis of the metrics used for MXAI methods' evaluation is performed. Finally, an extensive discussion regarding the current challenges and future research directions is provided.

INDEX TERMS Artificial intelligence, deep learning, evaluation, explanation, multimodal explainable artificial intelligence, neural networks.

I. INTRODUCTION

Over the last decade, humanity has witnessed unprecedented advancements in the field of Artificial Intelligence (AI), largely due to the emergence of the so-called Deep Learning (DL) paradigm that relies on the deployment of large-scale artificial neural networks and high-performing (GPU-enabled) computational infrastructures [1]. The introduced algorithms have been adopted in numerous application areas, leading to ground-breaking solutions and tremendous performance improvements. For example, DL has revolutionized the nature of research in the fields of computer vision, Natural Language Processing (NLP), audio analysis, self-driving cars and robotics [2], [3], [4], [5], [6], to name a few.

Although AI-enabled solutions have consistently resulted into remarkable outcomes across various tasks to which

The associate editor coordinating the review of this manuscript and approving it for publication was Valentina E. Balas.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

FIGURE 1. **Difference between unimodal and multimodal XAI: a) Unimodal explanation (saliency map) for image classification [7], and b) Multimodal explanation (visual and text) zero-shot learning [8].**

they have been applied, this is, however, accompanied by a cost in the exhibited transparency and trustworthiness of the developed systems [9]. In particular, it is typically very difficult to provide compact and precise explanations of the Neural Networks' (NNs) behavior and their eventual decision-making process. In order to address the latter challenge, the so-called eXplainable AI (XAI) research field has emerged, which aims, among others, at estimating meaningful explanations regarding the model's reasoning procedure [9], [10], [11], [12], [13], [14], [15], [16]. More specifically, how a prediction model works, how input data are used and what are the most critical/contributing features are some of the questions that XAI aims to answer. In this respect, XAI methods contribute towards making the models for the primary prediction task more transparent, while also significantly supporting the process of improving their performance. Indicative results of one of the most popular XAI methods are presented in Fig. 1a, where the Gradient-weighted Class Activation Mapping (Grad-CAM) approach is used for detecting and removing model bias in an image classification scenario [17].

While a large body of research works devoted to XAI has already been introduced, more recently the so-called Multimodal XAI (MXAI) approaches have been proposed, which naturally extend the fundamental principles and goals of unimodal XAI to the multimodal case (i.e. AI/DL-empowered methods that involve multiple types of modalities). The current work makes use of the term ***modality*** as defined in [18], which states that "***Modality refers to the way in which something happens or is experienced and a research problem is characterized as multimodal when it includes multiple such modalities***". In other words, the current work considers as MXAI methods those XAI approaches where: a) Multiple (two or more) modalities are used by the primary prediction model, b) Multiple (two or more) modalities are used for producing the primary models' behavior explanation, or c) The unimodal feature spaces of the primary model's input and the generated explanation are different (e.g. image/visual classification prediction (input) and textual explanation (output)). Therefore, MXAI refers to methods that overall involve multiple modalities in the primary and the explanation tasks. In this context, the most common modalities that are typically used in multimodal analytics schemes (and which in turn require multimodal analysis for explaining their behavior) include visual, audio, text, tabular, and graph data [8], [19], [20], [21].

Investigating the scope of MXAI in more detail, questions like "How?", "What?" and "Why?" need to be addressed in the multimodal setting, i.e. taking into account the increased AI/ML model complexity and the respective data multimodality, which inevitably requires a different mathematical consideration than that in the conventional unimodal XAI case. For example, it is evident that the two modalities that are used in Fig. 1b to explain a particular classification decision provide a more complete, justified and human-understandable explanation, compared to the respective one in Fig. 1a. Therefore, MXAI approaches can support explanations using multiple modalities that are complementary and cover more aspects of explainability. While aiming at developing robust MXAI schemes, the following key challenges (that also constitute clear advances and contradistinctions from the unimodal XAI case), need to be addressed (among others):

- Identification of the modality with the highest impact on the overall model's prediction;
- Discovery of the most salient features in each modality;
- Explanation of the employed modality fusion scheme itself, in order to understand the overall model's inference procedure and the cross-modal correlations that are learned;
- Detection of causal relations in the model's input-output data streams for generating explanations more easily conceivable by humans;
- Analysis of algorithms of increased complexity that are usually needed for handling multimodal data.

As clearly outlined by the above analysis, MXAI emerges as a rather challenging, yet promising, research field that has recently received increased attention, while its particular characteristics differentiate it significantly from the conventional unimodal XAI paradigm. In this context, this study

FIGURE 2. Key bibliometric analytics regarding MXAI literature: a) Article types, b) Top-15 most popular venues, and c) Scientific areas.

aims to comprehensively investigate, summarize and analyze in depth all recent advances and current research trends in the field of MXAI. Specifically, the main contributions of this work are:

- Formulation of a **comprehensive registry of AI/ML-boosted tasks** (i.e. specific application cases) where MXAI solutions have been applied until now, as well as a **thorough report of the relevant datasets** that have been used for learning/evaluating explanations in multimodal scenarios;
- **Systematic and comprehensive analysis/review of MXAI methods** that have been introduced so far, taking into account **three main criteria**:
  - the **number of involved modalities** (in the AI/ML module);
  - the **processing stage** at which explanations are generated;
  - the **type of the adopted methodology** (i.e. the actual mechanism and mathematical formalization) for producing explanations;
- Thorough examination of the **metrics used for MXAI methods' evaluation**;
- Extensive discussion of **current challenges** and **future research directions** in the field.

It needs to be highlighted that the current work significantly extends the respective/previous review of the MXAI methods' landscape presented in [32], since the present study: a) Follows a more systematic approach for analyzing the literature grounded on three different criteria (namely the number of involved modalities, the explanation generation stage and the type of the adopted methodology), b) Focuses on investigating only MXAI methods', while maintaining not to revisit the relevant multimodal processing and unimodal XAI fields, and c) Provides a thorough analysis of the metrics used for quantitative MXAI evaluation.

The manuscript is organized as follows: Section II outlines the adopted methodology for reviewing the MXAI literature. Section III discusses the various tasks and applications in which MXAI has been utilized, along with the respective publicly available datasets. Section IV systematically presents the methods for generating explanations in multimodal scenarios. Section V details the metrics used for evaluating explanations in the multimodal setting. Finally, Section VI discusses the current challenges and future research

TABLE 1. Datasets for MXAI learning and evaluation for a VQA task.

Dataset (date)	Input space	Output space	Description
VQA-X (2018) [22]	Image, textual questions	Textual answer, textual explanation	Extension of VQA 2.0 [23] (204,721 images, 1 million questions, 6.5 million of answers and 295,538 complementary QA pairs) with one textual explanation for each training QA pair, three for those in test/val sets and visual explanations collected from humans.
VQA-E (2018) [24]	Image, textual questions	Textual answer, textual explanation	VQA 2.0 extension with explanations created from image/question/answer triplets.
TextVQA-X (2021) [25]	Image, textual questions	Textual answer, visual & textual explanation	In addition to images (11,681 instances), questions (15,374 instances) and answers, it contains visual and textual explanations (67,055 instances each).
SCIENCEQA (2022) [26]	Image and/or textual context, textual questions	Textual answer, lecture, text explanation	Multiple choice questions in various scientific fields annotated with lectures (17,798 instances) and textual explanations (19,202 instances). Some questions have image context (10,332 instances), some text context (10,220 instances) and some both (6,532 instances).
CLEVR (2017) [27]	Image, textual questions, symbolic program	Textual answer	Image (100,000 instances), questions (864,968 instances) and answers (849,980) for training and validation. Only images and questions for testing. Annotated scene graphs and symbolic program representations for images in training and validation sets.
CLEVRCoGenT (2017) [27]	Image, textual questions, symbolic program	Textual answer	Same with CLEVR, but the objects and their colors in images are captured under two different conditions.
SHAPES (2016) [28]	Image, textual questions, symbolic program	Textual answer	Yes/no answers and annotated programs for each question (244 unique questions for 15,616 images).
Visual genome (2017) [29]	Image, textual questions, scene graph, attributes	Textual answer	It contains images (108,077 instances) with region descriptions (4,3 million instances), objects (1,4 million instances), relations between objects (1,5 million instances), attributes (1,6 million instances), scenes (108,249 instances), region graphs (3,8 million instances) and question-answers (1,773,258 instances).
VQA-HAT (2017) [30]	Image, textual questions	Textual answer, attention map	Extension of the VQA dataset of [31] (614,163 questions, 203,721 images and 10 answers for each train and validation question) with human attention maps for a subset of the training and the validation sets (58,475 and 1,374 instances, respectively).

directions in the field, while Section VII concludes the paper.

II. LITERATURE REVIEW METHODOLOGY

In order to efficiently and thoroughly identify/map the MXAI literature, while at the same time detecting key concepts and trends, a systematic approach was followed for ensuring comprehensiveness and relevance of the selected research works. In particular, a structured literature review methodology was adopted, consisting of the (iterative) main steps described below.

A. SCOPE AND OBJECTIVES FORMULATION

The fundamental goal of the performed survey study was to review the MXAI literature, i.e. AI/DL-empowered XAI methods that involve multiple types of modalities, according to the MXAI definition provided in Section I. In particular, the focus was on identifying the most recent advancements and works with substantial contribution in the field, emphasizing on the specific objectives of tracking: a) The utilized datasets, b) The main categories of methods based on multiple criteria, c) The employed evaluation protocols, and d) Current challenges and future research directions.

B. LITERATURE SEARCH

In order to ensure broad and thorough coverage of the relevant literature, the search strategy involved querying across several major scientific databases, including IEEE Xplore, Google Scholar, Scopus, DBLP, arXiv and Web of Science. The actual search was performed by combining targeted relevant keywords/terms (e.g. 'explainable artificial intelligence', 'multimodal', 'explainability', 'interpretability', etc.) and Boolean operators (i.e. 'AND', 'OR', 'NOT'). In order to guarantee contemporary relevance, search primarily focused on research works published within the last five years; however, certain seminal studies introduced earlier were also included. An example of the query used in the Scopus database is as follows:

TITLE-ABS-KEY("Multimodal Explainable Artificial Intelligence" OR "Multimodal XAI" OR "Explainable AI" OR "XAI" OR "Explainability" OR "Interpretability")
AND TITLE-ABS-KEY ("Methodology" OR "Framework" OR "Algorithm" OR "Approach" OR "Evaluation" OR "Challenges" OR "Future research directions")
AND TITLE-ABS-KEY ("Deep learning" OR "Neural networks" OR "Machine learning" OR "Natural language processing" OR "Computer vision" OR "Speech recognition" OR "Multimodal fusion" OR "Multimodal models")
AND PUBYEAR > 2020
AND PUBYEAR < 2025
AND (SRCTYPE (j) OR SRCTYPE (p))
AND EXCLUDE (DOCTYPE, "bk")

Multiple search steps were performed in an iterative way, involving refinements to the employed keywords, so as to

TABLE 2. Datasets for MXAI learning and evaluation for different primary prediction tasks.

Task	Dataset (date)	Input space	Output space	Description
Visual captioning	COCO-2017 (2017) [33]	Image	Caption, bbox, segmentation, labels	Object and stuff image segmentations from 80 and 91 categories, respectively, and 5 captions per image.
	COCO-2014 (2014) [33]	Image	Caption, bbox, segmentation, labels	Object and stuff image segmentations from 80 and 91 categories, respectively, and 5 captions per image.
	Flirck30K entities (2015) [34]	Image	Caption, bbox	244,000 co-reference chains for the Flickr30K captions [35].
	Visual genome (2017) [29]	Image, scene graph, attributes	Caption	See Table 1
Zero-shot learning & fine-grained classification	CUB (2011) [36]	Image, attributes	Labels, textual explanation, bbox	Bird images from 200 categories with 15 locations, 312 attributes, 1 bounding box and 5 sentences per image.
Activity recognition	Olympic sports (2010) [37]	Video	Labels, bbox, attributes	Videos from 16 sport categories with attributes and bounding boxes.
	UCF101 (2012) [38]	Video	Labels, bbox, attributes	Videos of 24 activities with attributes and bounding boxes.
	ACT-X (2018) [22]	Image	Labels, textual & visual explanation	Images of 397 activities with descriptions, 3 textual explanations and visual explanations per image.
Self-driving cars	BDD-X (2018) [39]	Video	Actions, textual explanations	77 hours of driving videos containing 3-4 actions each, annotated with descriptions and explanations.
Visual common-sense reasoning	VCR (2019) [40]	Image, multiple choice questions	Answer, textual rationale	290,000 multiple choice image questions, including a rationale for the correct answer.
Breast mass diagnosis	DDSM (1998) [41]	Image	Labels, textual explanation	605 mass images with BI-RADS descriptions and mass locations.
Outfit recommendations	ExpFashion (2019) [42]	Image	Recommendations, text explanations	200,745 outfit images with at least 3 explanations for each.

access more relevant works. Moreover, the list of references of each research article was also analyzed, in order to identify additional relevant studies.

C. SCREENING

Initial screening relied on excluding duplicate records, non-English papers and articles without full-text access, in order to maintain the integrity of the review study. Then, article selection was performed taking into account title and abstract information, so as to eliminate irrelevant works. Subsequently, in-depth full-text review was performed for considering only research studies that: a) Focus on MXAI techniques for producing explanations (i.e. excluding conventional unimodal XAI methods), and b) Exhibit substantial theoretical and/or experimental contributions. Eventually, a total of 94 articles were selected for analysis and were included as references in the current manuscript.

Key bibliometric analytics regarding the performed literature review study are illustrated in Fig. 2, while in-depth analysis of the identified MXAI works is provided in Sections III-VI.

III. PREDICTION TASKS AND DATASETS USED IN MXAI SCENARIOS

MXAI approaches are employed for explaining the behavior of AI/ML prediction models that involve multiple types of modalities and/or producing explanations (of the primary AI model's behavior) using different or multiple modalities. In this respect, the most popular AI/ML tasks that exhibit the above characteristics and are shown to benefit from the generation of MXAI explanations are: Visual Question Answering (VQA) [22], [23], [24], [25], [26], [43], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], [61], [62], [63], visual captioning [49], [50], [64], [65], [66], [67], [68], [69], [69], [70], visual common-sense reasoning (including also the visual dialog case extension) [40], [71], [72], recommendation systems [42], [73], fine-grained visual classification and zero-shot learning [8], [20], [74], [75], [76], [77], [78], [79], [80], [81], [82], (human) activity recognition [19], [22], [83], [84], emotion/sentiment recognition (including also hate speech detection and affect recognition) [21], [85], [86], [87], [88], [89], [90], [91], candidate screening [92], self-driving cars [39], [93], computer aided diagnosis [94], [95], [96], sleep range classification [97], and many more.

Regarding the main datasets that have been introduced so far for developing and evaluating MXAI approaches, these are illustrated in Tables 1 and 2. In particular, the tables group the various datasets with respect to the corresponding application task (as discussed above), while they also include information about the type of the input/output modalities,

FIGURE 3. Categorization of MXAI approaches: a) Main classes of conventional (unimodal) XAI methods that can also be adopted in MXAI analysis, b) Basic categories that can be considered only for MXAI approaches.

date and a short description for each entry. It needs to be mentioned that Table 1 includes only datasets related to the VQA task, due to the increased popularity of this particular application case. On the other hand, it must be highlighted that additional relevant datasets are publicly available for each task (e.g. VQA); however, the ones indicated in Tables 1 and 2 provide the framework for enabling MXAI evaluation (and not only the development of the related AI/ML module for the primary prediction task).

IV. EXPLAINABILITY IN THE MULTIMODAL SETTING

This section investigates in depth the application of XAI methods under multimodal scenarios. In particular, Section IV-A provides the fundamental criteria that are considered for classifying the various MXAI approaches, as well as the resulting categories. Subsequently, Section IV-B investigates conventional unimodal XAI approaches that are directly, or after being extended in a straightforward way, applied to multimodal cases. Then, Sections IV-C, IV-D and IV-E analyze the MXAI literature in relation to the number of the involved modalities, the processing stage at which explanations are generated and the adopted methodology, respectively.

A. CRITERIA AND RESULTING CATEGORIES OF MXAI METHODS

Aiming at systematically analyzing the MXAI literature, an initial categorization can be made based on the classes of methods that have been widely adopted for the unimodal XAI case [9], [12], [16], [32], [98], [99]. In particular, the following criteria, which are graphically illustrated in Fig. 3a, can be used:

- **Aim of explanation**: Depending on the scope of the explanation, MXAI methods can be considered as introspective [17] (i.e. focusing on the internal logic and behavior of the examined model) or justification-related [20], [32] (i.e. aiming at interpreting the model's predictions in a more human-friendly way).
- **Exclusiveness of explanation**: With respect to the extent of the validity of the generated explanation, MXAI methods can be either global [21] or local [100]. The former methods explain the overall model's behavior, while the latter focus on interpreting the produced decisions concerning specific data points.
- **Model dependency**: Concerning whether the internal structure and the architectural characteristics of the examined model are taken into account, MXAI approaches can be model-agnostic [100] (i.e. they can be applied to any type of ML model, regardless of its architecture or underlying algorithms) or model-specific [22] (i.e. they can be tailored to specific types/architectures of AI models).
- **Data dependency**: MXAI methods can be adapted so as to operate only for specific data types [99], e.g. visual, text, tabular, audio, etc., i.e. they can essentially be data type-specific approaches.

A critical characteristic of any MXAI method is the number of modalities that are being involved, considering both those that relate to the primary prediction task and those associated with the generated explanation. To this end, the following

FIGURE 4. MXAI categories with respect to the number of the involved modalities in the primary prediction model input and the generated explanation: a) Unimodal task and unimodal explanation (UU) [20], b) Unimodal task and multimodal explanation (UM) [8], c) Multimodal task and unimodal explanation (MU) [56], and d) Multimodal task and multimodal explanation (MM) [22].

four main categories can be identified (Fig. 3b), while representative examples of each category are also graphically illustrated in Fig. 4:

- **Unimodal task and unimodal explanation (UU)**: Comprises methods that for a unimodal primary prediction task estimate an explanation using a single, but different, modality. Fig. 4a illustrates an example of an image classification algorithm (primary task), associated with a textual description/justification of the produced prediction (generated explanation) [20].
- **Unimodal task and multimodal explanation (UM)**: Includes approaches that for a unimodal primary task produce an explanation consisting of at least two modalities. Fig. 4b depicts an example of an image classification model (primary task), accompanied by a saliency map and a textual rationale description of the estimated prediction (generated explanation) [8].
- **Multimodal task and unimodal explanation (MU)**: Comprises methods that for a multimodal primary task employing at least two modalities produce a unimodal explanation. Fig. 4c illustrates an example of a VQA approach that receives as input an image and an associated question in textual format (primary task), while an image saliency map is estimated for interpreting the model's output (generated explanation) [56].
- **Multimodal task and multimodal explanation (MM)**: Includes approaches that for a multimodal primary task (using at least two modalities) estimate also a multimodal explanation. Fig. 4d illustrates an example of a VQA method, which is associated with both a saliency map and a textual rationale justification (generated explanation) [22].

The development/processing stage (with respect to the model for the primary task) at which explanations are produced significantly differentiates the design of the corresponding MXAI approaches. Using the latter as a criterion, the following three main categories can be defined (in Fig. 3b):

- **Intrinsic**: Comprises methods that produce explanations by analyzing the internal structure and the parameters of the model that has been developed for the primary task [16], [32], [87].
- **Post-hoc**: Corresponds to those methods that do not investigate the internal architecture of the original model (developed for the primary task), i.e. approaches that are solely based on the analysis of the primary model's output [65], [69].
- **Separate module**: Incorporates methods that develop a distinct model (i.e. different from the one deployed for the primary task), in order to generate the required explanations [22], [100]. Depending on the actual phase during which the (new) explanation module is constructed, the approaches under this category can be further divided into:
  - -- Joint training, where the explanation module is trained along with the model used for the primary task [20], [80];

TABLE 3. Categorization of MXAI methods based on the number of involved modalities, explanation stage and adopted methodology.

Task & explanation modalities	Explanation stage	Methodology
UU	Joint training	Reasoning [20], [80], graph/example-based [101], graph/attribute/rule-based [84]
	Incremental training	Counterfactual [78]
UM	Post-hoc	Example-based [82], gradient/attribute-based [81], interactive/graph-based [83]
	Joint training	Example/attribute-based [74], gradient/attribute-based [8]
	Incremental training	Gradient/attribute-based [8], counterfactual [19], attention [39], counterfactual/reasoning [75], attribute/example-based [79]
MU	Intrinsic	Concept-based [21], fusion/graph-based [85], attention [49], [51]–[54], [56], [57], [62], [63], [66], [73], [89], [102], [103], attention/gradient-based [50], fusion/attention [88]
	Post-hoc	Clustering [86], ablation [91], [97], graph-based [55], concept-based [87], attention [69]
	Joint training	Reasoning [26], [48], attention [60], [64], example-based [23], interactive [72] , concept-based [70], graph-based [61]
	Incremental training	Reasoning [104], graph-based/reasoning [71], attention/attribute-based [67]
MM	Intrinsic	Attention [47], [105]
	Post-hoc	Gradient/occlusion [44], gradient/attribute-based [76], rule-based [92]
	Joint training	Attention/reasoning [22], [24], [93], attention [43], attention/interactive/graph-based [46], graph-based [25], reasoning [40]
	Incremental training	Attention [22], [45], [95], gradient/attribute-based [77]

-- Incremental training, where the explanation module is constructed after the model for the primary task has been developed [19], [45].

MXAI methods can also be classified based on commonly met methodologies (i.e. mathematical formalizations and mechanisms) that constitute fundamental building blocks of their processing pipeline. In particular, some of the proposed methodologies (e.g. causal-modeling [19], reasoning [20], graph-modeling [55] and attribute-based [8]) highly depend on the utilized data, while others (e.g. interactive [46], fusion [85] and attention-based [47]) are mostly related to the architecture of the examined model for the primary prediction task.

Table 3 summarizes the main MXAI approaches of the literature, which are hierarchically organized based on the number of the involved modalities, the explanation stage and the adopted methodology.

B. UNIMODAL XAI METHODS EXTENDED TO MULTIMODAL SCENARIOS

Several conventional XAI approaches, despite having been originally developed for unimodal tasks, can be extended to the multimodal setting in a relatively straightforward way. In the following, key representative examples of such methods are discussed in more detail.

1) DISENTANGLED MULTIMODAL EXPLANATIONS (DIME)

It constitutes a local model-agnostic method that extends the fundamental idea of the Local Interpretable Model-agnostic Explanations (LIME) [107] approach to multi-modal scenarios. In particular, DIME [100] disentangles the examined model into unimodal contributions and multimodal interactions, assuming that the overall model is formed as the aggregation of them. More specifically, LIME is applied separately to each unimodal contribution and to the multimodal interaction of the resulting aggregation.

2) GRADIENT-WEIGHTED CLASS ACTIVATION MAPPING (GRAD-CAM)

The fundamental conceptualization of the Class Activation Map (CAM) [108] approach is applicable to any model comprised of convolutional and a Global Average Pooling (GAP) layers. In this respect, the Grad-CAM [17] approach follows a back-propagation gradient-based scheme, where salient points in the input data that lead to the achieved prediction are identified, regardless of the number of the involved modalities.

3) GRAD-CAM++

It is a modified version of the original Grad-CAM method, which uses a weighted average of the positive gradients of the target class. The updated method is able to locate multiple occurrences of the same type of objects in an image more accurately. Grad-CAM++ [106] makes use of the same fundamental mechanism as Grad-CAM, which renders it possible to be applied to multimodal models as well, like those used for the tasks of VQA and image captioning (Fig. 5a).

4) LIFT-CAM

The conventional Grad-CAM method [17] can be extended to the multimodal setting, as a combination of the SHapley

FIGURE 5. Conventional XAI methods' generated explanations for multimodal scenarios: a) Grad-CAM++ for image captioning [106], b) LIFT-CAM for a VQA task [58], c) RISE for image captioning [68], and d) Guided-backpropagation for a VQA task [59].

Additive exPlanations (SHAP) [109] and Deep Learning Important FeaTures (DeepLIFT) [110] approaches. In particular, LIFT-CAM [58] determines activation maps, making use of SHAP values; DeepLIFT is used in this case as an approximation to SHAP coefficients, due to their intractable nature (Fig. 5b).

5) COMBINATION OF SHAP AND GRAD-CAM

A combined approach of the conventional Grad-CAM and the SHAP methods is investigated in [94], in order to provide explanations in a skin lesion diagnosis application. In particular, the SHAP method is used for handling tabular input data (e.g. age, gender, etc.), while Grad-CAM is employed for processing the visual input; hence, overall resulting in a multimodal explanation scheme.

6) INTEGRATED GRADIENTS (IG)

The original method [111] aims to correlate the model's prediction results with the input data and to identify the most salient features, by estimating the path from the prediction output to the original input, using gradient-related information. Extending to the multimodal case, IG is applied in a VQA scenario in [112], in order to identify which words in the question are significant for producing the estimated answer.

7) LAYERWISE RELEVANCE PROPAGATION (LRP)

The conventional LRP method relies on propagating the estimated prediction through the layers of the neural network [113]. Moving to the multimodal case, a variant of the LRP is used considering a convolutional neural network that receives as input three different sequences of MRI images, in order to detect characteristics of the brain that have an effect on its aging [114]. In a similar way, multimodal LRP is also employed in [65], in order to identify both the important pixels in the image and the contribution of previously generated words in the produced caption.

8) RANDOMIZED INPUT SAMPLING FOR EXPLANATION (RISE)

The conventional RISE method [68] makes use of masked image inputs, in order to subsequently observe the resulting effect on the produced model's class prediction. The weighted aggregation of the masks and the respective prediction scores create the saliency map used for explanation. RISE's extension to the multimodal setting is straightforward. For example, saliency maps are estimated for each word in the produced textual description in an image captioning task [68] (Fig. 5c).

9) GUIDED-BACKPROPAGATION

The original backpropagation method [115] aims at visualizing specific image features that have been detected from certain neurons, e.g. in an image analysis task. Its application to the multimodal setting is also relatively straightforward. For example, saliency maps are produced in a VQA task [59], where the Hadamard product of the visual and the textual features results in more accurate importance maps, compared to those resulted from using the respective attention weights [47], [116] (Fig. 5d).

10) CONCEPT ACTIVATION VECTORS (CAV)

The original method [117] aims at identifying specific semantic concepts, whose presence in the input data affects the model's prediction with respect to a given class. The latter relies on the use of differential/derivative estimations to assess how important is a certain pre-defined concept to a particular classification decision. In a multimodal scenario, CAV vectors can be used for explaining multimodal (audio,

FIGURE 6. UU MXAI methods' generated explanations: a) Textual [20], b) Textual [80], c) Object relationships [101], and d) Scene graph [84].

video and text data) emotion recognition, taking into account concepts that are relevant to different types of emotions [87].

C. MXAI CATEGORIZATION BASED ON THE NUMBER OF THE INVOLVED MODALITIES

MXAI methods can be classified taking into account the different combinations of the number of modalities involved in the primary prediction task and the generated explanation, as discussed in Section IV-A and graphically illustrated in Fig. 4.

1) UNIMODAL TASK AND UNIMODAL EXPLANATION (UU)

Methods belonging to this category often involve the visual modality as input for the primary prediction task, while the explanation can be either in textual format or in the form of a graph for efficiently representing correlations among entities. Fig. 6 illustrates indicative examples of representative literature approaches.

a: TEXTUAL EXPLANATIONS

Explanations of models' predictions in textual form provide an efficient way for intuitively detailing the actual models' reasoning process. As an example, in a bird image classification task, the justifications of object classification decisions are generated based on the key discriminator factors of different bird species in [20]; the method employs reinforcement learning techniques, while the generated explanations are class/prediction-specific, focusing on the explanation and not on a general-purpose image caption (Fig. 6a). Similarly, the activations of all neural network layers are concatenated in [80], targeting to provide a concrete explanation for the prediction of the bird class (Fig. 6b). Hendricks et al. [78] extend the latter idea to the case of estimating counterfactual explanations, where missing characteristics/properties of the depicted objects are identified, aiming at providing a rationale regarding why a specific bird does not belong to a given category. In a different image-based task, where the goal is to train a recommender system to also provide meaningful explanations, user comments are treated as ground-truth explanations in [42], in order to justify the matching of top and bottom clothes (e.g. shirts and trousers).

b: GRAPH-BASED EXPLANATIONS

Graph models provide an elegant way for formulating explanations, since they are inherently capable of representing multiple and diverse types of relationships among entities. Such an approach is particularly suitable for several application tasks requiring the justification of the prediction outcome based on the detected relationships, like in visual captioning, image classification and action recognition, to name a few. In [101], an object-relation graph is created for connecting the objects detected in an image with various relationships (Fig. 6c), where three types of connections between objects in the examined image are supported. Zhuo et al. [84] follow a graph-based approach in video-based action recognition, where for each input video frame the output consists of a scene graph depicting the relations among the objects inside the scenery (Fig. 6d).

2) UNIMODAL TASK AND MULTIMODAL EXPLANATION (UM)

Explanations of AI models' behavior are usually in a unimodal form, which, however, can sometimes lead to incomplete representation and understanding of the model's

FIGURE 7. UM MXAI methods' generated explanations: a) Heatmap-text [81], b) Heatmap-text [39], c) Example-attribute [74], and d) Graph-text [83].

reasoning process. To this end, multimodal explanations can often be advantageous, since they provide additional/supplementary explanatory statements in different modalities. In the following, UM MXAI approaches, which are grouped with respect to the combination of modalities involved in the generated explanation, are discussed in detail.

a: HEATMAP-TEXT EXPLANATION

A reasoning approach is followed in [81], where the primary task attribute prediction is initially performed, while a combination of the prediction's embeddings is subsequently used for producing the final classification. Then, using a back-propagation formalism, a score for each image attribute is obtained and the top-3 ones are subsequently utilized for forming a textual explanation, while Grad-CAM is also used for generating corresponding visual explanations (Fig. 7a). Similarly, a reasoning-based MXAI method focusing on the activities of acceleration and course prediction in the scenario of a self-driving car is investigated in [39], where an attention map provides introspective inference behind the obtained prediction results, while text is also generated as a supporting means to justify them. Moreover, the Grad-CAM approach is employed within the zero-shot learning paradigm for estimating a CAM (i.e. heatmap) for each image attribute in [8], where a whole image CAM representation is obtained by merging the individual CAMs of all considered attributes to form a heatmap for the examined class, while a textual description is also generated using visual features, attributes and latent embeddings.

b: IMAGE-TEXT EXPLANATION

Natural language explanations do not always provide a sufficient and complete justification of a model's particular decision; hence, grounding the produced explanations also on the visual modality is shown to be beneficial. In this respect, the methods of [20] and [75] generate textual explanations for the examined image, chunk them into phrases and estimate a score for each of them, indicating image relevance; the highest-scoring chunks are subsequently used for reinforcing the textual explanations (Fig. 7b). In [79], the textual part of the explanation is combined with complementary visual examples (e.g. images), which are associated with the generated textual justification; the estimated example can belong to the same or an opposing semantic class. Moreover, the textual justification is accompanied by an additional counterfactual part in [19], where action predictions in videos are justified using a combination of textual explanations and a bounding box of the predicted and the opposing classes.

c: EXAMPLE-ATTRIBUTE EXPLANATION

Aiming at increasing the expressiveness and the completeness of the generated explanations, estimated attributes (e.g.

FIGURE 8. MU MXAI methods' generated explanations: a) Attention map for image captioning [66], b) Textual explanation for a VQA task using graphs [55], c) Example-based combined with attention mechanism for a VQA task [56], and d) Concept-based mechanism for an interpretable emotion recognition task [21].

shape, color, etc.) can be combined with exemplary instances of the original input image data. In particular, Ul Hassan et al. [74] follow a visual search approach, apart from incorporating attributes for producing a justification of the model's prediction, in a visual classification task (Fig. 7c). This visual search strategy relies on exploiting the convolutional features of the utilized NN model, in order to retrieve similar data instances that serve as additional justification information. Moreover, an adversarial approach is followed in [82], in order to estimate complementary/counter-examples, each associated with individual image attributes (e.g. bird bill shape, color, etc.).

d: GRAPH-TEXT EXPLANATION

The expressiveness of graphs in representing accurate explanations can be further reinforced, by combining them with an interaction mechanism. In particular, a video is provided as input to a network that outputs a graph, based on the detected objects and actions [83]. Apart from the explanation composed of the identified semantic concepts and their relations, the graph is utilized by an interactive question-answering agent that can answer inquiries regarding the graph structure and the video (Fig. 7d); hence, producing additional textual explanations or alternative graph explanations.

3) MULTIMODAL TASK AND UNIMODAL EXPLANATION (MU)

MXAI techniques are particularly suitable for addressing explainability needs concerning multimodal prediction tasks, i.e. when the examined AI model receives as input data from multiple modalities. In the following, MU MXAI methods that produce unimodal explanations are discussed. For the sake of clarity, the presentation is organized according to the particular methodology/mechanism that is adopted for producing the explanations.

a: ATTENTION-BASED

Attention schemes constitute the most widely used mechanism in MU MXAI methods, due to their inherent ability to adjust the analysis focus (during the explanation generation process) on the important parts of the input data (e.g. critical image regions and words). One of the most common primary tasks that attention mechanisms exhibit wide applicability is that of visual captioning. In particular, a stochastic and a deterministic attention mechanism are used in [66], in order to estimate heatmaps of the image regions that correspond to the predicted captions (Fig. 8a). Similarly, spatial and spatiotemporal saliency maps are computed for the estimated image and video captions in [69], respectively. Han and Choi [64] compute relevance scores between the detected objects and the respective words in the estimated captions, aiming at explaining why certain words are generated. Moreover, attended regions in the visual medium are employed in captioning and VQA settings [49], in order to interpret the produced captions or answers, respectively.

Attention-based MU MXAI methods are also widely used in VQA applications, where the typical approach consists

of estimating an attention map over the input image that highlights important regions that mostly contribute to the generated visual answer. Specifically, a pooling operator is used twice in the examined model in [102], where the attended regions are used for grounding the produced answer. Additionally, Zhu et al. [52] combine an attention map with a bounding box of the detected object for estimating a visual explanation. Multiple attention layers are used in [53] and various attention distributions over the image are considered in [54], in order to localize the important image regions. Moreover, objects and regions of attention are used during the training phase of a visual grounding model in [63]. Alipour et al. [51] introduce an explainable VQA system, in order to examine the impact of the explanations on the users in terms of model competency. Furthermore, important visual regions are estimated within a clothing recommendation system [73], analyzing both product images and customer reviews.

Among other application cases of MU MXAI methods that rely on the use of attention mechanisms, object detection and human-generated heatmaps are used by the so-called Human Importance-aware Network Tuning (HINT) [50] approach, in order to improve performance in vision-language models, by leveraging gradient-based explanations. Additionally, the most contributing words in hate speech detection are estimated in [89], using text, social and demographic features. Moreover, attention to both image and text is employed in [103], in order to visualize regions of interest that correspond to a specific hypothesis in a visual entailment task.

b: GRAPH-MODELLING

Graph models are suitable for producing unimodal explanations from a multimodal input feature space. In particular, bounding boxes are used to interpret the answer regarding the counting abilities (e.g. counting objects in images) of a model in [60], while a text explanation is estimated using an image scene graph and an attention map in [55] (Fig. 8b). Additionally, Vedantam et al. [61] follow a probabilistic approach to create an interpretable VQA model, using symbolic programs that can track the model's reasoning. Symbolic programs are also used in [62], in order to generate visualizations of each step of the model's inference procedure.

c: REASONING

An emerging line of research in the field of MXAI concerns the use of various reasoning formalisms, which often aim to mimic the human way of inference for generating improved explanations. In that respect, textual explanations provided by humans are used for creating a VQA model in [48], which learns not only to estimate answers but also makes use of retrieved explanations to generate more accurate ones. Additionally, Lu et al. [26] produce textual explanations in response to scientific questions, making use of lectures (related to the subject in question), complementary to the use of relevant images and questions. Pre-trained transformer-based language models exploit objects and their relations to provide full-sentence answers and rationales in a VQA setting in [71]. Moreover, a visual dialog approach is presented in [72], where a conversation regarding the image at hand is produced, by answering consecutive questions about it.

d: EXAMPLE-BASED

An alternative approach, aiming at increasing the expressiveness of the generated explanations, consists of the use of exemplary data instances. In this context, counter-visual instances are used in a VQA scenario in [23], targeting to both support the quality of the produced answers and to restrict the model's bias (with respect to the input questions), by collecting similar images for which for the same question a different answer is estimated. Additionally, images that support/oppose the produced answer in a VQA setting are used to create a map in [56], which depicts the image regions that humans would have focused on (Fig. 8c).

e: CONCEPT-BASED

Explanations that are grounded on concept-based representations are often shown to be advantageous, regarding their interpretation by the human user. In this respect, semantic information, extracted using Latent Dirichlet Analysis (LDA), is considered for detecting activations that correspond to certain topics (e.g. people, dancing and eating, among others) in a video captioning setting in [70]; the explanation is eventually provided by a numerical metric that quantifies the correlation of a given neuron activation with a topic in a video frame. Additionally, Tsai et al. [21] estimate local and global explanations for multimodal sentiment analysis using a routing approach to identify the importance of data (either characterizing individual modalities or cross-modal features), based on explainable hidden embeddings (Fig. 8d).

f: FUSION

Fusion schemes are also shown to be beneficial in the explanation generation process. In particular, an explainable fusion scheme is proposed in [85], which is applied in the context of sentiment analysis and provides a numerical effectiveness metric calculated from the fusion parameters, indicating how modality interactions contribute to the final prediction. Additionally, scores illustrating modality contributions (for two different representations of each modality) are used for an emotion recognition task in [88]. Furthermore, Wu et al. [90] propose the use of capsules to obtain estimates on modality dynamics indicative of their contribution to the final emotion recognition prediction.

g: ABLATION

Ablation-based approaches can also be efficient in generating meaningful explanations. Specifically, the overall significance score of each modality is estimated in a medical signal classification application in [97]. Additionally, Lin et al. [91] introduce a feature importance method (for obtaining

FIGURE 9. MM MXAI methods' generated explanations: a) Image- and text-heatmap [47], b) Image heatmap and text explanation [22], c) Image heatmap and text attributes [76], and d) Personality trait scores and text justification [92].

significance scores at the sensor level) and an ablation approach (for estimating feature importance at the signal level) in a multimodal affect recognition setting in [91].

h: CLUSTERING

Clustering techniques are also shown suitable in MXAI applications. In [86], the embeddings of each NN layer are used to demonstrate that deep layers distinguish emotions better than shallow layers in a multimodal emotion recognition model, utilizing audio and text features.

4) MULTIMODAL TASK AND MULTIMODAL EXPLANATION (MM)

MM approaches constitute the most complex type, in terms of the number of modalities in the input and output feature spaces, since they support multimodal information both for the primary prediction task and the generated explanation. In the following, different types of MM MXAI methods are discussed, which are grouped according to the (most common) combinations of modalities/representations in the generated explanations.

a: IMAGE HEATMAP AND TEXT EXPLANATION

The so far most popular MM MXAI category of methods relies on the use of a heatmap for the input image, along with supplementary text for generating a more complete explanation. In particular, an attention map and textual justifications are estimated for the VQA and the action recognition tasks in [22] (Fig. 9b). Similarly, Li et al. [24] employ an attention mechanism for estimating important image regions and corresponding textual rationales are produced for complementing the explanation output for a VQA application. Interactive explanation schemes that allow users to interact with the model, when it provides erroneous answers, and subsequently to improve its behavior are employed in [46], in order to investigate their efficiency in a VQA setting. Additionally, Patro et al. [43] adopt a correlation estimation approach between answers and explanations, in order to increase the robustness of the examined model, but also to provide meaningful explanations for the produced answer. Ground-truth visual and textual explanations are used in [25], in order to train a model to generate corresponding multimodal explanations. Moreover, attention maps over video frames and respective textual justifications, regarding control predictions in a self-driving vehicle setting, are presented in [93].

b: IMAGE- AND TEXT-HEATMAP EXPLANATION

One of the most common techniques relies on the combination of heatmaps produced (separately) for both the image and the text modality. In particular, backpropagation and occlusion methods are used in [44], in order to identify words (in questions) and image areas of significant importance when answering a visual question. Additionally, Lu et al. [47] use attention schemes at multiple levels (between image and text phrases information streams) in a VQA application, in order to produce heatmaps over the input image and the respective question [47] (Fig. 9a). Along with the textual answer in a VQA scenario, a justification is also produced in [45], based on the segmented image regions that correspond to specific words in the estimated rationale.

c: IMAGE HEATMAP AND TEXT ATTRIBUTES

Apart from combining image heatmaps with textual explanations, extracted text attributes can also be incorporated. In particular, domain-specific knowledge is employed in [76], following an attribute-based formalism for the text stream and a Grad-CAM-grounded approach for the visual one

FIGURE 10. Intrinsic MXAI methods' generated explanations: a) Attention heatmap [53], b) Image and attention heatmap [50], and c) Modality importance [85].

in a zero-shot learning approach (Fig. 9c). Similarly, a gradient-based method for zero-shot learning and fine-grained classification is presented in [77], where, apart from the produced visual heatmap images, a text explanation generator is developed that takes into account the detected attributes (e.g. color, shape, etc.).

d: OTHER MULTIMODAL EXPLANATIONS

The wide set of combinations that can be considered regarding the input modalities for the primary prediction task and the corresponding ones used for generating explanations allows for multiple and significantly diverse MM MXAI techniques to be developed. In particular, in an attempt to move towards cognition-level understanding, the so-called Recognition to Cognition Networks (R2C) are introduced in [40], where, given a challenging question about an image, the networks target to answer correctly and then to provide a rationale justifying their answer; the explanations receive the form of bounding boxes that are associated with semantic concepts in the generated rationale. Additionally, Cao et al. [105] investigate the information learned in visual-language transformer models, aiming at identifying which is the most significant modality that captures the cross-modal interactions from certain attention heads, as well as the hidden visual or linguistic information that is stored in the latter. Moreover, an image with the aligned face of a candidate, normalized personality trait scores and a respective rationale are produced in [92], in order to explain the outcome of an interview in a job screening scenario (Fig. 9d).

D. MXAI CATEGORIZATION BASED ON THE EXPLANATION STAGE

MXAI methods can also be grouped into different categories based on the development/deployment stage (with respect to the primary prediction task model) at which explanations are learned/produced, as discussed in Section IV-A and detailed in the remainder of this section.

1) INTRINSIC

The methods of this category make use of the primary prediction model's parameters to estimate meaningful interpretations of the produced results. In the following, intrinsic MXAI approaches are discussed in detail, while they are organized based on the most common types of generated explanations.

- **Attention heatmap**: Importance masks are efficient for demonstrating the components of the input data that are critical for explaining the model's inference process. In particular, a so-called Multimodal Compact Bilinear (MCB) pooling operator is introduced in [102], in order to address the problem of intractable outer product calculation between matrices in VQA settings; the operator is used twice for creating spatial attention maps over the image to justify the estimated answer. Yang et al. [53] introduce the so-called Stacked Attention Networks (SANs) (Fig. 10a), which follow a multi-step reasoning approach; each layer identifies individual parts of the input image that the SAN has attended to produce answers for a VQA task. Additionally, a stochastic and a deterministic approach are presented in [66], in order to compute the attention weights for an image captioning model and to estimate the corresponding importance maps for each generated word. In order to force the model to focus on the same image regions that a human would do, an example-based approach is followed in [56], where a nearest-neighbor method is adopted concerning the semantic

FIGURE 11. Post-hoc MXAI methods' generated explanations: a) Video captioning [69], b) Fine-grained visual classification [81], c) VQA [44], and d) Biomedical signal processing [97].

similarity of images and distances between attention weights of similar examples are maintained to be lower than those of counter-examples. Patro et al. [57] adopt a gradient-based approach for computing the loss gradients and attention maps are subsequently produced for the answers in a VQA setting, along with the corresponding uncertainties in the prediction process. Bidirectional Encoder Representations from Transformers (BERT) attention maps are investigated in a VQA application in [51], which are shown to concentrate on the most relevant areas in the examined image. Moreover, attention schemes are used to decompose the model's reasoning into individual/consecutive steps that converge to the final generated answer in a VQA problem in [62]. Furthermore, Chen et al. [73] employ attention schemes in recommender systems, in order to detect important regions in product images that significantly affect the recommendation decision, taking into account user/item properties and the corresponding image. Text, tabular and graph data are considered by a self-attention mechanism in [89], aiming at estimating salient features in hate-speech detection.

- **Image and attention heatmap**: In order to estimate more complete and informative explanations, attention heatmaps have also been combined with parts of the input image. In particular, an LSTM model is introduced in the context of a VQA task in [52], where a spatial attention mechanism relates words in the provided questions with corresponding image regions and eventually estimates bounding boxes of critical detected objects (over an attention heatmap) to ground the produced answer. Similarly, attention weight parameters are considered in [49], in order to estimate salient regions that relate to each word in the generated captions and answers in a VQA task. Additionally, Zhang et al. [63] make use of an attention-based model within a supervised setting in a VQA task, in order to estimate region/object-based groundings in the examined image. Selvaraju et al. [50] follow a Grad-CAM-oriented approach for obtaining salient image regions and associating them with saliency scores (Fig. 10b). Moreover, a self-attention mechanism is used for both images and text in a visual entailment task in [103], in order to generate either heatmap- or image-based explanations.
- **Modality importance**: A critical aspect in MXAI analysis comprises the assessment of the level of importance of the various modalities involved in the primary prediction task. In this respect, a graph-based fusion approach is implemented in [85], in order to dynamically model the interactions between modalities in the emotion recognition and sentiment analysis tasks (Fig. 10c). Additionally, Tsai et al. [21] estimate local and global explanations for emotion recognition, where dynamically adjusted importance measures are assigned to unimodal and bimodal interactions for each processed data sample. Moreover, a so-called capsule network is integrated into a routing mechanism in [90], in order to estimate a contribution score for each modality involved in an emotion recognition task. An attention-based scheme is also used in [88] to align text and speech representations in an emotion recognition setting; eventually, the extent of contribution of each modality to the final outcome is calculated. Furthermore, Cao et al. [105] make use of the learned attention weights and heads of visual-language models, in order to visualize what type of information has been encoded in each head, which modality is more important and the interactions of the latter.

2) POST-HOC

MXAI approaches under this category aim to produce meaningful explanations of a model's behavior after the AI prediction module has been applied and its results are made available, i.e. the model for the primary prediction task is considered as a 'black box' one. In the following, post-hoc

MXAI approaches are presented in detail, while they are grouped taking into account the relevant (and most popular) primary prediction tasks.

- **Visual classification and captioning**: An attention mechanism is employed to provide saliency maps (both spatial and temporal) for each generated word for the tasks of image and video captioning in [69]; the mechanism takes into account the decrease observed in the output word probabilities when frames or image regions are removed from the input (Fig. 11a). Additionally, an inherently explainable decision tree is employed in [92], in order to interpret positive/negative outcomes, when deciding about interview invitations. A zero-shot learning approach for image-level classification is introduced in [76], which targets the modeling of semantically meaningful concepts that have been implicitly learned by individual neurons in convolutional neural networks and produces a heatmap and the top activated attributes. Moreover, a back-propagation-based approach is adopted in [81], in order to identify the extent of the detected attributes' (e.g. color, shape, etc.) contribution to the final outcome, while also producing complementary saliency maps and textual justifications (Fig. 11b). Gulshad and Smeulders [82] select examples from an opposing semantic class while considering visual attributes extracted by a model trained in an adversarial way; the explanation is eventually formed by combining attributes/examples of the original and the counter class. Furthermore, a graph representation is produced in [83], incorporating concepts detected in the input video; then, an interactive agent is applied on top to provide explanations of the depicted action.
- **Visual question answering**: A guided backpropagation method is combined with a part occlusion one (applied to segments of input images and questions) in [44], in order to identify visually important image regions and to predict the contribution of individual question words in the VQA generated answer (Fig. 11c). Additionally, Ghosh et al. [55] employ an attention-based heatmap to define the most relevant parts in an image, while a scene graph is also incorporated, along with NLP technologies, to provide natural language rationales, using the detected entities and their relations.
- **Biomedical signal processing**: A random forest algorithm is adapted to a multimodal affect recognition setting in [91], in order to provide sensor-level feature importance assessments and signal-level explanations. Additionally, principal components of layered embeddings of audio and text features are considered in [86], in order to reason about the separation of the detected emotion classes. Ellis et al. [97] apply an ablation-oriented analysis at the modality level, in order to determine the significance of different signals in sleep stage classification, by replacing one of the captured modalities with noisy data and observing the variation in the prediction probabilities (Fig. 11d).

3) SEPARATE MODULE

This category comprises methods that rely on the development of a new/distinct module that generates explanations of the primary model's behavior.

a: JOINT TRAINING

Under this consideration, the models for the primary prediction and the explanation generation tasks are constructed together, allowing in this way the training process of each module to affect the respective procedure of the other one. In the following, joint training MXAI approaches are discussed, while being grouped according to the primary task of concern.

- **Visual question answering**: An explanation module is trained together with the actual VQA model in [23], in order to provide a similar image with a different answer (for a given question), essentially forming an example-based explanation. An attention-based mechanism (for the visual information stream) and an LSTM architecture (for the textual information stream) are combined on top of the actual model for producing the required explanations in [22]; the models can be trained either jointly or incrementally, depending on the selected task (namely VQA or action recognition, respectively). Additionally, a multi-task learning architecture is introduced in [24], where a computational model is trained to generate an explanation, along with the answer predicted by the primary model (Fig. 12a). On the other hand, symbolic programs, which are used to represent the model's reasoning, are utilized in [61], in order to justify the generated answers. Patro et al. [43] make use of a generator module (for producing textual explanations) and a respective correlation one (for ensuring that the produced answer complies with the estimated explanation). Moreover, an interactive approach is followed in [46], where an attention-based scene graph, which takes into account the detected objects, is incorporated. Explanations are used in a competitive fashion to improve both the answers and the corresponding rationales in [48]. Furthermore, Rao et al. [25] make use of ground-truth visual and textual explanations for learning to interpret the generated answer.
- **Visual classification**: An image classifier is jointly trained with a respective textual explanation generation module in [20], following a reinforcement learning approach. Similarly, a reasoning-based method is developed in [80], in order to rationalize classification decisions, using the activations of the fully connected layers of the classification model as input to the explanation generation one. Additionally, ul Hassan et al. [74] introduce an LSTM-based explanation module equipped with two attention mechanisms (one applied to the considered attributes and one operating on top of the target categories), in order to produce rationales for the generated prediction and also to (supplementarily)

FIGURE 12. Separate module MXAI methods' generated explanations: a) Joint training for VQA [24], b) Incremental and joint training for zero-shot learning [8], c) Incremental training for VQA [45], and d) Incremental training for video classification [19].

retrieve semantically similar images. A saliency map and an associated attribute-based textual justification are used in [8] (Fig. 12b), where the joint training strategy is shown to lead to significant performance improvements, compared to the separate training case. Moreover, Zhuo et al. [84] exploit scene graph-based representations (making use of objects, their attributes and object relations information), aiming at both tracking the illustrated actions across the video frames and demonstrating the model's reasoning process.

- **Visual captioning**: Topics extracted using LDA analysis are used in [70], in order to jointly train a model with both a negative log-likelihood (for the primary captioning task) and an interpretable loss (for measuring the feature agreement with respect to semantically similar topics). Additionally, Han and Choi [64] design a module for examining whether the captioning model has focused on relevant image objects, by measuring the compliance of the detected objects/concepts and the correspondingly generated words using an attention mechanism.
- **Other**: Lu et al. [101] concentrate on detecting relationships among image objects and formulating a respective explanation graph, trained in an end-to-end fashion. Textual rationales are generated for answering follow-up questions regarding the examined image in a visual dialog setting in [72]. Additionally, responses to visual questions and textual rationales are jointly produced in the context of a visual common-sense reasoning task in [40]. Kim et al. [93] introduce a combined approach for simultaneously realising next action prediction and justification in a self-driving cars application case. Moreover, recommendation and explanation functionalities are jointly developed in a reasoning framework in [42], aiming at providing explanations for each suggested item.

b: INCREMENTAL TRAINING

In many cases, the explanation generation module is constructed separately from the respective primary task one (namely after the primary AI model is developed), due to factors like reduction in required training computations, lack of availability of explanations in the training data, etc. In the following, incremental training MXAI methods are presented that are grouped with respect to the primary task of concern.

- **Visual question answering**: Park et al. [22] estimate both a textual rationale and a visual attention heatmap, by jointly or incrementally training the explanation module (depending on the task at hand). Additionally, Li et al. [104] initially extract attributes (e.g. sit, phone, bench, etc.) and generate textual explanations for the image, while a reasoning module subsequently utilizes the extracted justifications to infer an answer to the given question. Attention mechanisms are used over segmented images (Fig. 12c) in [45], in order to produce explanations consistent with the predicted answers. A transformer-based model is employed for providing justifications in various vision-language tasks in [71], making use of pre-trained GPT-2 language models.
- **Visual classification**: A so-called Deep Multimodal Explanation (DME) model, which incorporates a joint visual-attribute embedding module and a multi-channel explanation one trained in an end-to-end fashion, is introduced in [8] (Fig. 12b), addressing the needs of explainable zero-shot learning applications. An explainable AI agent is developed in [75], where a phrase-critic model receives as input an image and a candidate explanation, and outputs a score indicating how good the candidate explanation is grounded on the image. Additionally, a method to generate textual counterfactual explanations is presented in [78], focusing on inspecting which evidence data in the input is missing, but which could potentially contribute to a different

classification decision if present in the examined image. Kanehira et al. [19] exploit a spatio-temporal video region (tube) and textual attributes (e.g. using a pole, flipping, etc.) for estimating counterfactual multimodal explanations (Fig. 12d). Linguistic explanations and a set of visual examples for rendering the classification decision interpretable are used in [79], where explanations are parameterized by three different NNs (namely a predictor, a linguistic explainer and an example selector model). Moreover, Wickramanayake et al. [77] generate post-hoc linguistic justifications to rationalize the decision of a CNN, where a decision-relevance metric that measures the faithfulness of an explanation to a model's reasoning is used. Furthermore, Lee et al. [95] introduce an explanation module that can be stacked to any Computer-Aided Diagnosis (CADx) network model, in order to provide rationales for medical diagnosis application cases.

- **Other**: Kim et al. [39] produce introspective explanations in a self-driving vehicles scenario, which combine a visual (spatial) attention model (that identifies image regions that potentially influence the model's output) and an attention-based video-to-text module (that produces textual explanations of the primary model's actions). Additionally, Fang et al. [67] introduce textually-grounded explanations in computer vision applications, where the text descriptions are decomposed into three levels (namely entity, semantic attribute and color information) for progressively realizing compositional justification.

E. MXAI CATEGORIZATION BASED ON THE ADOPTED METHODOLOGY

Within the field of MXAI research, there are certain commonly met methodologies (i.e. mathematical formalizations and mechanisms) that constitute fundamental building blocks of the respective proposed methods, as discussed in Section IV-A. To this end, using the particularly adopted methodology as a classification criterion, MXAI approaches can be categorized as follows:

- **Casual-modeling**: It mainly corresponds to techniques that support explainability through the use of counterfactual examples. In particular, counterfactual explanations aim at determining which are the minimum allowed modifications to the input data, so as to alter the primary model's prediction to a specific/pre-defined, yet different, output [19], [78].
- **Reasoning**: Similarly to numerous approaches in various AI-boosted application domains, extensions of the traditional multi-task learning conceptualization are also applied to the MXAI field. In particular, the examined AI model is jointly trained so as, apart from providing only predictions for the primary task, also to produce explanations about the generated outcomes and the overall model's reasoning process [20], [22], [39].
- **Graph-modeling**: Analysis based on graphs constitutes a particularly valuable approach in MXAI scenarios, since they are highly efficient in identifying correlations among data points (also of diverse nature), while they also enable the generation of insightful representations/visualizations of the detected relations. In particular, graphical models have been extensively used in computer vision applications, where, for example, scene graphs have been employed for representing an image and for subsequently estimating answers/explanations for a given question [46]. Similarly, graphs have been utilized for estimating symbolic representations of textual sources, targeting, for example, to produce a step-by-step representation of the AI model's inference process [62].
- **Attribute-based**: Identifying attributes in the input data, which are critical for shaping the model's predictions, constitutes an efficient way for subsequently producing accurate explanations of the model's behavior. For the case of visual input data, attributes can refer to certain image characteristics (e.g. color and shape of objects), aiming at eventually estimating an overall image saliency map [8]. In a similar fashion, natural language rationales can be formed to explain individual model's decisions, e.g. determining the specific attributes/arguments behind a bird being classified as belonging to a particular breed [74].
- **Interactive**: Approaches under this category allow humans to interactively intervene in the explanation generation process. Specifically, the user is allowed to provide feedback about the model's produced predictions and, subsequently, the model exploits the collected information for improving its performance or posing questions about the assessment of the results. For example, humans are capable of evaluating AI models' predictions and corresponding explanations through an interactive framework in [46], while follow-up questions are presented to the user for justification purposes in [72].
- **Fusion-based**: Numerous and diverse fusion schemes have been adopted in multimodal prediction models for estimating comprehensive data representations and achieving accurate prediction results. For example, a multiplication operator of image and question representations is typically used in VQA approaches. In this context, investigating the way that fusion is applied and the effects that the latter mechanism imposes can provide significant insights in terms of explainability purposes, e.g. identifying how each modality contributes to the produced outcome [88].
- **Attention-based**: The scope of the attention mechanism is to enable the trained model to focus only on specific features in the input data space that are important for realizing robust predictions, e.g. specific image patches or individual words in textual phrases. The latter is achieved by estimating weights that modulate

the model's attention on the input data. The estimated computed weights can be utilized though for providing meaningful explanations of the model's decisions, e.g. in the form of visualizations of the attended regions in an image or words in a sentence [22], [47], [49], [57].

V. EVALUATION OF GENERATED EXPLANATIONS

Evaluation schemes in XAI applications aim at assessing the accuracy and efficiency of the generated explanations, i.e. the ability to which the XAI method at hand explains the behavior/decisions of the model developed for the primary prediction task. The main categories to which XAI (including MXAI techniques) evaluation approaches can be roughly classified to comprise [118]: a) Application-grounded, i.e. implementing experimental evaluations with domain experts, b) Human-grounded, i.e. conducting experiments with humans, without necessarily being domain experts in the examined field, and c) Functionally-grounded, i.e. using a formal definition of explainability in order to assess the quality of the produced explanations. With respect to the particular case of MXAI methods, the evaluation protocol can consider/assess the involved modalities independently (with textual and visual explanations being the most common ones) or also take into account cross-modal interrelations/dependencies.

A. EVALUATION OF TEXTUAL EXPLANATIONS

In order to assess the quality (i.e. accuracy, correctness, etc.) of the generated textual explanations (that are usually in the form of natural language sentences), a large body of the relevant literature relies on the implementation of user-centered studies, i.e. experimental procedures that require the provision of human-user assessments (e.g. in the form of questionnaires, explanation gradations/rankings, etc.) [20], [22], [39], [48], [71], [75]. The latter is mainly due to: a) Lack of publicly available datasets that include ground-truth explanation-related annotations, and b) Inherent difficulty in determining precisely and unequivocally what a 'good' explanation is for a given data input. However, when ground-truth explanation data are available, the following main metrics (where BLEU-1,2,3,4, ROUGE, METEOR, CIDEr and SPICE have been borrowed from the NLP literature) have been widely used, in order to evaluate the similarity between the generated textual justification and the corresponding target/ground-truth one:

- **BLEU**: It counts an n-gram (i.e. continuous sequence of n words) based matching score between the candidate explanation and the ground-truth one, regardless of the word order [8], [22], [24], [25], [26], [43], [45], [48], [74], [80], [93], [95], [104].
- **ROUGE**: Among the various individual metrics belonging to this category, ROUGE-L takes into account the longest common sub-sequence between the reference and the predicted explanation phrase [8], [22], [24], [25], [26], [43], [45], [48], [74], [80], [93], [95], [104].
- **METEOR**: While originally designed to address shortcomings of the respective BLEU score, its estimation is based on an explicit word-to-word matching between the generated explanation and all respective reference ones [8], [20], [22], [24], [25], [39], [43], [45], [48], [74], [80], [93], [104].
- **CIDEr**: It assesses the degree of consensus between a candidate explanation sentence and a set of reference ones, by examining how often n-grams in the candidate phrase appear in the reference ones [8], [20], [22], [24], [25], [39], [43], [45], [48], [74], [80], [93], [95], [104].
- **SPICE**: It is based on the creation of semantic scene graphs from dependency parse trees created using the candidate and reference explanation sentences, taking into account objects, attributes and their relationships; it relies on assessing the generated explanation/sentence quality using an F-score metric calculated over tuples (conjunction of logical propositions) belonging to both graphs [22], [43], [45], [48], [93].
- **Ratio of unique or novel sentences**: These focus on identifying/counting the explanation sentences that have not been generated before and those that do not exist in the training set [119].
- **Cosine-similarity**: It measures the correspondence of the generated natural language explanations and the ground-truth ones, taking into account the respective available semantic embeddings [120].
- **Phrase error and accuracy with counterfactual test**: These metrics are applicable to the case of counterfactual textual explanations [78]. In particular, phrase error estimates the degree of similarity of the counterfactual and the ground-truth sentence, i.e. an ideal score should be equal to zero. Additionally, accuracy aims at measuring the decrease in class prediction performance, where additional counterfactual text is provided (along with the originally generated textual explanation) as input to the classifier.
- **Class similarity**: It relies on estimating the relevance of the generated explanation for a particular semantic class with the ground-truth ones for the same class [20], by employing the CIDEr metric between the generated explanations and all reference sentences of the examined class (and not just the ground-truth ones for the specific image in question).
- **Class ranking**: It is an extension of the class similarity concept to further validate whether the generated explanations are class relevant. It relies on the estimation of the similarity score for each examined sentence and every considered class [20].
- **Relevance of visual attributes**: Such metrics measure the proportion of accurately represented ground truth attributes in the top-k generated explanations for each instance [76] or use the estimated attributes to assess the relevance of the generated explanations with the visual features that the examined model has learned [77]. The latter is accomplished by counting the words that are

common in the ground-truth description of an object and the generated justification, and then calculating the ratio of this count to the total number of words in the ground-truth description.

- **Relevance of textual attributes**: It evaluates the accuracy of the generated explanations (i.e. predicted attributes (in text form) and estimated captions), by utilizing a cosine similarity-based formalism.
- **Word importance**: It is based on estimating the importance of the individual words comprising the generated explanations under different evaluation scenarios, e.g. removing the top-3 relevant words [65], consideration of part-of-speech tags [44], etc.
- **Position of first and number of relevant explanations**: These estimate the ranking position of the first relevant explanation in the set of all generated ones and the number of relevant explanations in the top-5 produced ones [55].

B. EVALUATION OF VISUAL EXPLANATIONS

Similarly to the case of evaluating textual justifications, a large portion of the MXAI literature that produces visual explanations relies on the implementation of user-centered studies. For the latter case, qualitative assessment of the produced outcomes is usually realized by superimposing the explanation (e.g. heatmap) to the input visual data (either image or video) and, subsequently, the human-user is involved, so as to judge whether the identified/highlighted areas correspond to truly important pixels/points for the primary model's prediction. On the other hand, when relevant ground-truth information is available, quantitative evaluation can be performed by: a) Estimating the correspondence of the generated explanations (e.g. in the form of target bounding-boxes, human attention and segmentation maps, etc.) to the ground-truth ones, and b) Examining the contribution of individual image pixels and regions to the model decisions, by observing the effect of removing or superimposing them from/to the model's input data. When ground-truth explanation-related information is provided, the following metrics/approaches are commonly used in the literature:

- **Intersection-over-Union (IoU)**: This measures the correspondence between a given reference mask or bounding-box and the respective one associated with the explanation [8], [25].
- **Mask alignment**: It evaluates the visual alignment between the ground-truth and the generated explanation masks, focusing on e.g. examining whether the maximum attention weights reside inside the ground-truth bounding-box area [52].
- **Average fraction of activations**: It estimates the average percentage of the points (belonging to the produced activation maps) that lie within the provided ground-truth bounding boxes [17].
- **Earth Mover's Distance (EMD)**: It measures the similarity/distance between the produced attention/explanation map and the human-provided one, while considering them as two different probability distributions [22], [45], [57].
- **Rank correlation**: It relies initially on the ranking of the pixels of the produced/available heatmaps according to their spatial attention values and, subsequently, estimating the correlation between these ordered lists of pixels [22], [30], [43], [44], [50], [56], [57].
- **Foreground Attention Rate (FAR)**: It measures the degree of compliance of the detected foreground objects with the provided attention heatmap over the examined image [81].
- **Quality of attention/saliency map**: It assesses the accuracy of the generated heatmaps, by estimating the degree of consistency with the respective human-annotated ground-truth ones [121].
- **Precision and recall**: The conventional precision and recall performance metrics are extended to the case of evaluating the generated visual explanations, by e.g. estimating how often the center of the generated attention map overlaps with the available ground-truth region annotation [62].
- **Attention correctness** [122]: Specifically designed for attention models used in visual captioning tasks [65], [69], it is equal to the normalized sum of all attention weights that relate to a given word that, at the same time, correspond to the respective ground-truth annotated regions.
- **Negative class accuracy**: It aims at assessing the validity of the produced counterfactual explanation, by examining the extent to which the output explanation changes when the estimated attended region is removed [19].

Assessing the explanation evaluation issue from a causal relation modeling perspective, the analysis focuses on examining whether the generated visual regions contribute to the correct prediction and how changes in the input affect the final prediction result, i.e. not being constrained only to the case of examining solely the alignment of visual explanations with ground-truth data. In this context, the following measures/approaches are widely used:

- **Deletion**: It identifies critical pixels, which, when removed from an image, lead to a drop in detection performance for a given semantic class [68].
- **Insertion**: It determines important pixels, which, when added to an image, lead to an increase in the detection performance for the examined semantic class [68].
- **Average drop**: It corresponds to the average decrease in the model's prediction performance for a given class, when only explanation-related information is provided as input to the model, instead of the full original image [106].
- **Average drop in deletion**: It is similar to the 'average drop' metric described above, but it makes use of the prediction score for the class of interest, when considering the inverted explanation heatmap [58].

- **Increase in confidence**: It corresponds to the increase in the model's prediction confidence, when providing only explanation-related information as input to the model, instead of the original image [106].
- **Win percentage**: It compares the performance of different methods, by estimating the frequency that the performance achieved by a given method, when using only the generated explanation heatmap, is higher than that obtained by other/competitive approaches [106].

C. MULTIMODAL EVALUATION

While the wide majority of the literature approaches considers individual unimodal metrics for evaluating the quality of the generated MXAI explanations, methods that implement multimodal evaluation schemes (i.e. assessment approaches that take into account the correlations among the examined modalities) have also been proposed, as follows:

- **Concept accuracy** [19]: This aims at evaluating multimodal (visual and textual) counterfactual explanations for video classification. The metric estimates the compatibility of the estimated words (i.e. assigned attributes, e.g. using pole, flipping, etc.) in the textual justification and their visual counterparts, by comparing the conventional IoU metric of a given attribute-bounding-box pair with the respective pairs corresponding to the remaining generated counterfactual explanations.
- **Complementarity** [79]: It estimates the degree of complementarity between the visual and the textual part of an explanation, by utilizing the "reasoner"'s (it receives as input the generated explanation pairs) score for the class identified by the "predictor" (the actual prediction module) for each candidate combination of text and image explanation pairs.
- **Fidelity** [79]: This metric aims at examining how class predictions are obtained, by utilizing the generated explanations. The "reasoner" (it receives as input the generated explanation pairs) and the "predictor" (the actual prediction module) are compared in terms of accuracy and prediction consistency.

VI. CURRENT CHALLENGES AND FUTURE RESEARCH DIRECTIONS

Despite the large body of MXAI works that have recently been introduced, significant challenges and open research problems are still present, which if sufficiently addressed will further increase the efficiency and acceptance of MXAI schemes. It needs to be mentioned that the research directions described below are often applicable also to the unimodal XAI field.

A. CONVERGENCE TO FORMAL AND WIDELY ACCEPTED DEFINITIONS/TERMINOLOGY

Although many research studies have recently appeared in the MXAI field, little to no formality has been adopted, concerning the definitions and terminology used. In particular, many researchers make use of ad-hoc descriptions to delineate their research activities, while they often define 'explainability' and 'interpretability' in various (and often conflicting) ways. As a result, no concrete and widely accepted terminology is currently present. Defining what an explanation is and how its accuracy/efficiency can be measured using well-defined qualitative/quantitative norms and experimental frameworks, apart from enhancing formalization aspects in the field, will also significantly facilitate the comparative evaluation of the numerous proposed explainability methods [123]. The latter will also greatly assist in addressing current controversies, like assigning different terms to similar methods or associating similar names with fundamentally different (algorithmic) concepts.

B. USAGE OF ATTENTION MECHANISMS IN EXPLANATION SCHEMES

Attention mechanisms, apart from being used in numerous data analysis tasks, have also been utilized for generating explanations of the prediction models' behavior, typically in the form of visualization schemes (indicating words or image areas where the primary model focuses on) or feature importance metrics. However, several concerns and conflicting arguments have emerged, raising fundamental doubts regarding the suitability of attention mechanisms to produce actual explanations. In particular, experimental studies show that distributions between learned attention weights and gradient-based feature relevance methods are not highly correlated for similar predictions [124]; hence, conventional attention-based explanations can not be considered equivalent to others. However, contradictory experimental results move in the opposite direction, i.e. the usage of attention schemes for explanation generation is not always applicable, but it depends on the actual definition of explanation that is adopted in the particular application at hand [125]. In this context, more detailed and in-depth studies need to be conducted, in order to shed more light on whether and under which exact conditions attention schemes can be used for providing meaningful explanations, as well as how such methods relate to other non attention-based MXAI approaches.

C. GENERALIZATION ABILITY OF MXAI METHODS

The wide majority of the available methods has only been designed for specific AI model architectures, regarding the primary prediction task. For example, there is a significant number of methods that have been designed for the VQA scenario; however, such approaches have not been evaluated in other vision-language applications. Naturally, it can be well admitted that introducing model-specific explanation schemes is very restrictive and expensive. Robustly extending existing methods to other tasks and architectures would significantly save research resources and would likely lead to performance improvements.

D. EXTENSION OF MXAI SCHEMES TO MORE THAN TWO EXPLANATION MODALITIES

Most MXAI methods focus on producing unimodal or bimodal explanations. However, extending these representations to higher dimensionality multimodal feature spaces

(i.e. feature spaces that are composed of more than two modalities) would inevitably further increase the expressiveness and accuracy of the produced explanations.

E. ESTIMATION OF CAUSAL EXPLANATIONS

So far no significant attention has been given to the causality perspective of explanations, while causal relationships are the particular type of relations that humans inherently perceive. In this respect, causal explanations can enable the interpretation of how one event can lead to another one and, hence, the development of a deeper understanding of the world. On the other hand, identifying the factors that cause an event to occur can also facilitate the prediction of how the event might unfold and/or how it could/should be treated in the future [123]. Therefore, apart from identifying which features are important for a given model, how predictions are affected by modifications in the feature values is important to understand the model's reasoning process itself. In the context of the multimodal setting, causality needs to be examined in terms of how each individual modality and the correspondingly particular features affect the prediction outcome (and not simply identifying which features are important).

F. REMOVING BIAS IN TEXTUAL EXPLANATIONS

The main paradigm being followed for estimating textual explanations consists of collecting natural language rationales from humans and subsequently developing/training an explanation module with these descriptions as ground truth data. However, human textual annotations (especially when it comes to long textual justifications) typically contain (contradictory) biases that are related to the particular background and temperament of each involved individual. To this end, developing modulation schemes for identifying/removing bias and resolving conflicting annotation cases would significantly improve the quality of the generated textual annotations.

G. LACK OF GROUND-TRUTH VISUAL EXPLANATIONS

Contrary to the case of textual explanations, a lack of corresponding ground-truth visual explanation data, which would enable the development of explanation generation modules trained under a supervised learning scenario, is observed. Despite the fact that collecting such manual annotations can be very expensive, the availability of such data would greatly boost the development of more accurate and robust MXAI schemes.

H. INSUFFICIENT MXAI EVALUATION

One of the main gaps in MXAI research concerns the lack of standard evaluation metrics, protocols and benchmarks for assessing the quality of the produced explanations. This naturally hinders the objective comparative evaluation of different methods and the identification of the most efficient practices for generating accurate explanations. Additionally, the most widely adopted norm consists of examining each modality separately (i.e. not investigating the correlations among the different information streams of the produced explanations); hence, often leading to inaccurate observations and possible misconceptions regarding the obtained results. Moreover, the inherent human subjectivity, when it comes to explanation assessment, adds to the problem difficulty. Addressing all the aforementioned challenges, towards achieving objective MXAI evaluation results, should be coupled with the definition of suitable multimodal interpretability evaluation metrics, where only very few and highly task/model-specific ones have been introduced so far [19], [79].

I. EXPLANATIONS TARGETING SPECIFIC DOMAINS AND END-USERS

A critical aspect of explainability is that of producing explanations that are tailored to specific end-users, considering their individual needs and diverse backgrounds. Interpretation capabilities should be inherently tied to the nature of the particular user's experience and expertise (something that is usually neglected), and in the majority of the scenarios the produced explanations cannot be well understood by non-experts. The design of explanation schemes should consider the diverse backgrounds, knowledge and cognitive capabilities of the end-user interacting with a particular AI system. Users with limited technical expertise may require conceptually simpler explanations that would support a clear understanding of the system's decision-making process. On the other hand, users with domain-specific expertise may require more detailed and accurate information. On the other hand, tailoring explanations to different user profiles would further enhance the user's experience, by fostering comprehension, building trust and enabling effective collaboration between humans and AI systems.

VII. CONCLUSION

In this paper, a comprehensive, systematic and in depth study regarding the developments in field of Multimodal eXplainable Artificial Intelligence (MXAI) was presented. Initially, an extensive analysis of the relevant primary prediction tasks (e.g. image/video captioning, visual question answering, etc.) and the corresponding publicly available datasets, where MXAI approaches have been applied so far, was provided. Subsequently, a thorough and structured presentation of the MXAI methods of the literature was given, based on the following key criteria: a) The number of the involved modalities (regarding both the primary prediction model input and the generated explanation feature spaces), b) The development/deployment stage (with respect to the primary prediction task model) at which explanations are learned/produced, and c) The type of the methodology (i.e. mathematical formalism/mechanism) adopted for producing the actual explanations. Then, a detailed discussion regarding the issue of MXAI methods' evaluation was provided, emphasizing on outlining the relevant quantitative performance metrics. Finally, a comprehensive analysis of current challenges and future research directions in the field was given.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] A. Boukerche and Z. Hou, "Object detection using deep learning methods in traffic scenarios," *ACM Comput. Surveys*, vol. 54, no. 2, pp. 1–35, Mar. 2022.

[3] H. Du, H. Shi, D. Zeng, X.-P. Zhang, and T. Mei, "The elements of end-to-end deep face recognition: A survey of recent advances," *ACM Comput. Surveys*, vol. 54, no. 10s, pp. 1–42, Jan. 2022.

[4] S. Thermos, G. Th. Papadopoulos, P. Daras, and G. Potamianos, "Deep affordance-grounded sensorimotor object recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 49–57.

[5] S. Maruf, F. Saleh, and G. Haffari, "A survey on document-level neural machine translation: Methods and evaluation," *ACM Comput. Surveys*, vol. 54, no. 2, pp. 1–36, Mar. 2022.

[6] G. Th. Papadopoulos, M. Antona, and C. Stephanidis, "Towards open and expandable cognitive AI architectures for large-scale multi-agent human–robot collaborative learning," *IEEE Access*, vol. 9, pp. 73890–73909, 2021.

[7] R. R Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-CAM: Why did you say that?" 2016, *arXiv:1611.07450*.

[8] Y. Liu and T. Tuytelaars, "A deep multi-modal explanation model for zero-shot learning," *IEEE Trans. Image Process.*, vol. 29, pp. 4788–4803, 2020.

[9] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.

[10] E. S. Ortigossa, T. Goncalves, and L. G. Nonato, "Xplainable artificial intelligence (XAI)-from theory to methods and applications," *IEEE Access*, vol. 12, pp. 80799–80846, 2024.

[11] N. Tiwary, S. A. Mohd Noah, F. Fauzi, and T. S. Yee, "A review of explainable recommender systems utilizing knowledge graphs and reinforcement learning," *IEEE Access*, vol. 12, pp. 91999–92019, 2024.

[12] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surveys*, vol. 51, no. 5, pp. 1–42, Aug. 2018, doi: 10.1145/3236009.

[13] A. Hoenig, K. Roy, Y. T. Acquaah, S. Yi, and S. S. Desai, "Explainable AI for cyber-physical systems: Issues and challenges," *IEEE Access*, vol. 12, pp. 73113–73140, 2024.

[14] N. Y. Murad, M. H. Hasan, M. H. Azam, N. Yousuf, and J. S. Yalli, "Unraveling the black box: A review of explainable deep learning healthcare techniques," *IEEE Access*, vol. 12, pp. 66556–66568, 2024.

[15] T. Martins, A. M. de Almeida, E. Cardoso, and L. Nunes, "Explainable artificial intelligence (XAI): A systematic literature review on taxonomies and applications in finance," *IEEE Access*, vol. 12, pp. 618–629, 2024.

[16] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[17] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[18] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[19] A. Kanehira, K. Takemoto, S. Inayoshi, and T. Harada, "Multimodal explanations by predicting counterfactuality in videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8586–8594.

[20] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell, "Generating visual explanations," in *Computer Vision—ECCV*. Cham, Switzerland: Springer, 2016, pp. 3–19.

[21] Y.-H.-H. Tsai, M. Ma, M. Yang, R. Salakhutdinov, and L.-P. Morency, "Multimodal routing: Improving local and global interpretability of multimodal language analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 1823–1833.

[22] D. H. Park, L. A. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, and M. Rohrbach, "Multimodal explanations: Justifying decisions and pointing to the evidence," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8779–8788.

[23] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in VQA matter: Elevating the role of image understanding in visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6325–6334.

[24] Q. Li, Q. Tao, S. Joty, J. Cai, and J. Luo, "VQA-E: Explaining, elaborating, and enhancing your answers for visual questions," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 552–567.

[25] V. Nagaraj Rao, X. Zhen, K. Hovsepian, and M. Shen, "A first look: Towards explainable TextVQA models via visual and textual explanations," in *Proc. 3rd Workshop Multimodal Artif. Intell.*, 2021, pp. 19–29.

[26] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan, "Learn to explain: Multimodal reasoning via thought chains for science question answering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 1–20.

[27] J. Johnson, B. Hariharan, L. Van Der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 2901–2910.

[28] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Neural module networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Apr. 2016, pp. 39–48.

[29] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, May 2017.

[30] A. Das, H. Agrawal, L. Zitnick, D. Parikh, and D. Batra, "Human attention in visual question answering: Do humans and deep networks look at the same regions?" *Comput. Vis. Image Understand.*, vol. 163, pp. 90–100, Oct. 2017.

[31] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis.*, May 2015, pp. 2425–2433.

[32] G. Joshi, R. Walambe, and K. Kotecha, "A review on explainability in multimodal deep neural nets," *IEEE Access*, vol. 9, pp. 59800–59821, 2021.

[33] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, vol. 14, 2014, pp. 740–755.

[34] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer Image-to-Sentence models," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 74–93, May 2017.

[35] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 67–78, Dec. 2014.

[36] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-UCSD birds-200–2011 dataset," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. 2010-001, 2011.

[37] J. C. Niebles, C.-W. Chen, and L. Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 392–405.

[38] K. Soomro, A. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*.

[39] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata, "Textual explanations for self-driving vehicles," in *Computer Vision—ECCV*. Cham, Switzerland: Springer, 2018, pp. 577–593.

[40] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi, "From recognition to cognition: Visual commonsense reasoning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6713–6724.

[41] M. Heath, K. Bowyer, D. Kopans, P. Kegelmeyer, R. Moore, K. Chang, and S. Munishkumaran, "Current status of the digital database for screening mammography," in *Computational Imaging and Vision*. The Netherlands: Springer, 1998, pp. 457–460.

[42] Y. Lin, P. Ren, Z. Chen, Z. Ren, J. Ma, and M. de Rijke, "Explainable outfit recommendation with joint outfit matching and comment generation," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 8, pp. 1502–1516, Aug. 2020.

[43] B. N. Patro, S. Patel, and V. P. Namboodiri, "Robust explanations for visual question answering," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1566–1575.

[44] Y. Goyal, A. Mohapatra, D. Parikh, and D. Batra, "Towards transparent AI systems: Interpreting visual question answering models," *Comput. Vis. Pattern Recognit.*, vol. 1, pp. 1–2, Aug. 2016.

[45] J. Wu and R. Mooney, "Faithful multimodal explanation for visual question answering," in *Proc. ACL Workshop BlackboxNLP*, Aug. 2019, pp. 103–112.

[46] K. Alipour, J. P. Schulze, Y. Yao, A. Ziskind, and G. Burachas, "A study on multimodal and interactive explanations for visual question answering," 2020, *arXiv:2003.00431*.

[47] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 289–297.

[48] J. Wu, L. Chen, and R. J. Mooney, "Improving VQA and its explanations by comparing competing explanations," 2020, *arXiv:2006.15631*.

[49] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6077–6086.

[50] R. R. Selvaraju, S. Lee, Y. Shen, H. Jin, S. Ghosh, L. Heck, D. Batra, and D. Parikh, "Taking a HINT: Leveraging explanations to make vision and language models more grounded," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2591–2600.

[51] K. Alipour, A. Ray, X. Lin, J. P. Schulze, Y. Yao, and G. T. Burachas, "The impact of explanations on AI competency prediction in VQA," in *Proc. IEEE Int. Conf. Humanized Comput. Commun. Artif. Intell. (HCCAI)*, Apr. 2020, pp. 25–32.

[52] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, "Visual7W: Grounded question answering in images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4995–5004.

[53] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Apr. 2016, pp. 21–29.

[54] V. Kazemi and A. Elqursh, "Show, ask, attend, and answer: A strong baseline for visual question answering," 2017, *arXiv:1704.03162*.

[55] S. Ghosh, G. Burachas, A. Ray, and A. Ziskind, "Generating natural language explanations for visual question answering using scene graphs and visual attention," 2019, *arXiv:1902.05715*.

[56] B. Patro and V. P. Namboodiri, "Differential attention for visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7680–7688.

[57] B. Patro, M. Lunayach, S. Patel, and V. Namboodiri, "U-CAM: Visual explanation using uncertainty based class activation maps," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7443–7452.

[58] H. Jung and Y. Oh, "Towards better explanations of class activation mapping," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 1316–1324.

[59] J.-H. Kim and B.-T. Zhang, "Visual explanations from Hadamard product in multimodal deep networks," *ArXiv:1712.06228*, 2017.

[60] A. Trott, C. Xiong, and R. Socher, "Interpretable counting for visual question answering," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–20.

[61] R. Vedantam, K. Desai, S. Lee, M. Rohrbach, D. Batra, and D. Parikh, "Probabilistic neural symbolic models for interpretable visual question answering," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6428–6437.

[62] D. Mascharka, P. Tran, R. Soklaski, and A. Majumdar, "Transparency by design: Closing the gap between performance and interpretability in visual reasoning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4942–4950.

[63] Y. Zhang, J. C. Niebles, and A. Soto, "Interpretable visual question answering by visual grounding from attention supervision mining," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2019, pp. 349–357.

[64] S.-H. Han and H.-J. Choi, "Explainable image caption generator using attention and Bayesian inference," in *Proc. Int. Conf. Comput. Sci. Comput. Intell. (CSCI)*, 2018, pp. 478–481.

[65] J. Sun, S. Lapuschkin, W. Samek, and A. Binder, "Explain and improve: LRP-inference fine-tuning for image captioning models," 2020, *arXiv:2001.01037*.

[66] K. Xu, "Show, attend and tell: Neural image caption generation with visual attention," *Comput. Sci.*, vol. 2015, pp. 2048–2057, Feb. 2015.

[67] Z. Fang, S. Kong, C. Fowlkes, and Y. Yang, "Modularized textual grounding for counterfactual resilience," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6371–6381.

[68] V. Petsiuk, A. Das, and K. Saenko, "Rise: Randomized input sampling for explanation of black-box models," in *Proc. Brit. Mach. Vis. Conf.*, 2018, pp. 1–26.

[69] V. Ramanishka, A. Das, J. Zhang, and K. Saenko, "Top-down visual saliency guided by captions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, May 2017, pp. 7206–7215.

[70] Y. Dong, H. Su, J. Zhu, and B. Zhang, "Improving interpretability of deep neural networks with semantic information," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 4306–4314.

[71] A. Marasovic, C. Bhagavatula, J. S. Park, R. L. Bras, N. A. Smith, and Y. Choi, "Natural language rationales with full-stack visual reasoning: From pixels to semantic frames to commonsense graphs," in *Proc. EMNLP Finding*, 2020, pp. 2810–2829.

[72] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, S. Lee, J. M. F. Moura, D. Parikh, and D. Batra, "Visual dialog," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 5, pp. 1242–1256, May 2019.

[73] X. Chen, H. Chen, H. Xu, Y. Zhang, Y. Cao, Z. Qin, and H. Zha, "Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2019, pp. 765–774.

[74] M. ul Hassan, P. Mulhem, D. Pellerin, and G. Quénot, "Explaining visual classification using attributes," in *Proc. Int. Conf. Content-Based Multimedia Indexing (CBMI)*, Sep. 2019, pp. 1–6.

[75] L. A. Hendricks, R. Hu, T. Darrell, and Z. Akata, "Grounding visual explanations," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 264–279.

[76] R. R. Selvaraju, P. Chattopadhyay, M. Elhoseiny, T. Sharma, D. Batra, D. Parikh, and S. Lee, "Choose your neuron: Incorporating domain knowledge through neuron-importance," in *Computer Vision—ECCV*. Cham, Switzerland: Springer, 2018, pp. 540–556.

[77] S. Wickramanayake, W. Hsu, and M. L. Lee, "Flex: Faithful linguistic explanations for neural net based model decisions," in *Proc. AAAI Conf. Artif. Intell.*, Jul. 2019, pp. 2539–2546.

[78] L. Anne Hendricks, R. Hu, T. Darrell, and Z. Akata, "Generating counterfactual explanations with natural language," 2018, *arXiv:1806.09809*.

[79] A. Kanehira and T. Harada, "Learning to explain with complemental examples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8595–8603.

[80] S. Barratt, "InterpNET: Neural introspection for interpretable deep learning," 2017, *arXiv:1710.09511*.

[81] W. Xu, J. Wang, Y. Wang, G. Xu, D. Lin, W. Dai, and Y. Wu, "Where is the model looking at? – concentrate and explain the network attention," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 3, pp. 506–516, Mar. 2020.

[82] S. Gulshad and A. Smeulders, "Explaining with counter visual attributes and examples," in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2020, pp. 35–43, doi: 10.1145/3372278.3390672.

[83] S. N. Aakur, F. D. de Souza, and S. Sarkar, "An inherently explainable model for video activity interpretation," in *Proc. Workshops 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–24.

[84] T. Zhuo, Z. Cheng, P. Zhang, Y. Wong, and M. Kankanhalli, "Explainable video action reasoning via prior knowledge and state transitions," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 521–529.

[85] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 1–20.

[86] P. Kumar, V. Kaushik, and B. Raman, "Towards the explainability of multimodal speech emotion recognition," in *Proc. Interspeech*, Aug. 2021, pp. 1748–1752.

[87] A. R. Asokan, N. Kumar, A. V. Ragam, and S. S. Shylaja, "Interpretability for multimodal emotion recognition using concept activation vectors," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2022, pp. 1–8.

[88] P. Liu, K. Li, and H. Meng, "Group gated fusion on attention-based bidirectional alignment for multimodal emotion recognition," 2022, *arXiv:2201.06309*.

[89] P. Vijayaraghavan, H. Larochelle, and D. Roy, "Interpretable multi-modal hate speech detection," 2021, *arXiv:2103.01616*.

[90] J. Wu, S. Mai, and H. Hu, "Interpretable multimodal capsule fusion," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, no. 2, pp. 1815–1826, May 2022.

[91] J. Lin, S. Pan, C. S. Lee, and S. Oviatt, "An explainable deep fusion network for affect recognition using physiological signals," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, Apr. 2019, pp. 2069–2072.
[92] H. Kaya, F. Gürpinar, and A. A. Salah, "Multi-modal score fusion and decision trees for explainable automatic job candidate screening from video CVs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1651–1659.
[93] J. Kim, S. Moon, A. Rohrbach, T. Darrell, and J. Canny, "Advisable learning for self-driving vehicles by internalizing observation-to-action rules," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Apr. 2020, pp. 9658–9667.
[94] S. Wang, Y. Yin, D. Wang, Y. Wang, and Y. Jin, "Interpretability-based multimodal convolutional neural networks for skin lesion diagnosis," *IEEE Trans. Cybern.*, vol. 52, no. 12, pp. 12623–12637, Dec. 2022.
[95] H. Lee, S. T. Kim, and Y. M. Ro, "Generation of multimodal justification using visual word constraint model for explainable computer-aided diagnosis," in *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2019, pp. 21–29.
[96] A. Lucieri, M. N. Bajwa, S. A. Braun, M. I. Malik, A. Dengel, and S. Ahmed, "ExAID: A multimodal explanation framework for computer-aided diagnosis of skin lesions," *Comput. Methods Programs Biomed.*, vol. 215, Mar. 2022, Art. no. 106620.
[97] C. A. Ellis, R. Zhang, V. D. Calhoun, D. A. Carbajal, M. S. E. Sendi, M. D. Wang, and R. L. Miller, "A novel local ablation approach for explaining multimodal classifiers," in *Proc. IEEE 21st Int. Conf. Bioinf. Bioengineering (BIBE)*, Oct. 2021, pp. 1–6.
[98] F. Hohman, M. Kahng, R. Pienta, and D. H. Chau, "Visual analytics in deep learning: An interrogative survey for the next frontiers," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 8, pp. 2674–2693, Aug. 2019.
[99] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, 2020.
[100] Y. Lyu, P. P. Liang, Z. Deng, R. Salakhutdinov, and L.-P. Morency, "DIME: Fine-grained interpretations of multimodal models via disentangled local explanations," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, Jul. 2022, pp. 455–467, doi: 10.1145/3514094.3534148.
[101] C. Lu, R. Krishna, M. S. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 852–869.
[102] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 1–25.
[103] N. Xie, F. Lai, D. Doran, and A. Kadav, "Visual entailment: A novel task for fine-grained image understanding," 2019, *arXiv:1901.06706*.
[104] Q. Li, J. Fu, D. Yu, T. Mei, and J. Luo, "Tell-and-answer: Towards explainable visual question answering using attributes and captions," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 1338–1346.
[105] J. Cao, Z. Gan, Y. Cheng, L. Yu, Y.-C. Chen, and J. Liu, "Behind the scene: Revealing the secrets of pre-trained vision-and-language models," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 565–580.
[106] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized Gradient-Based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 839–847.
[107] M. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in *Proc. 2016 Conf. North American Chapter Association Computational Linguistics, Demonstrations*, Jun. 2016, pp. 97–101.
[108] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.
[109] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 4768–4777.
[110] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 3145–3153.
[111] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 3319–3328.
[112] P. K. Mudrakarta, A. Taly, M. Sundararajan, and K. Dhamdhere, "Did the model understand the question?" in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 1896–1906.
[113] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS One*, vol. 10, no. 7, Jul. 2015, Art. no. e0130140.
[114] S. M. Hofmann, F. Beyer, S. Lapuschkin, O. Goltermann, M. Loeffler, K.-R. Müller, A. Villringer, W. Samek, and A. V. Witte, "Towards the interpretability of deep learning models for multi-modal neuroimaging: Finding structural changes of the ageing brain," *NeuroImage*, vol. 261, Nov. 2022, Art. no. 119504.
[115] J. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," in *ICLR Workshop Track*, 2015.
[116] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 299–307.
[117] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, and F. Viegas, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2668–2677.
[118] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017, *arXiv:1702.08608*.
[119] Y. Wang, Z. Lin, X. Shen, S. Cohen, and G. W. Cottrell, "Skeleton key: Image captioning by skeleton-attribute decomposition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7378–7387.
[120] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 3982–3992.
[121] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," *Int. J. Comput. Vis.*, vol. 126, no. 10, pp. 1084–1102, Oct. 2018.
[122] C. Liu, J. Mao, F. Sha, and A. Yuille, "Attention correctness in neural image captioning," in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2017, vol. 31, no. 1, pp. 1–20.
[123] W. Saeed and C. Omlin, "Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities," *Knowl.-Based Syst.*, vol. 263, Mar. 2023, Art. no. 110273.
[124] S. Jain and B. C. Wallace, "Attention is not explanation," in *North American Chapter of the Association for Computational Linguistics*, 2019.
[125] S. Wiegreffe and Y. Pinter, "Attention is not not explanation," in *Proc. 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 11–20.

NIKOLAOS RODIS received the B.Sc. degree in mathematics from the University of Patras, Greece, in 2017, and the M.Sc. degree in statistics from Athens University of Economics and Business, Greece, in 2020. He is currently pursuing the Ph.D. degree with the Harokopio University of Athens, Greece. His current research interests include computer vision, deep learning, and explainable artificial intelligence.

CHRISTOS SARDIANOS received the B.Sc. degree in electronic engineering, the M.Sc. degree (Hons.) in informatics and telematics in the area of web engineering, and the Ph.D. degree in recommender systems. He is currently a Research Associate and Postdoctoral Researcher with the Department of Informatics and Telematics, Harokopio University of Athens. His main research interests include recommender systems and their applications, machine learning, data mining, and data analysis and management. He has authored more than 45 articles published in international peer-reviewed journals and conferences on recommender systems, knowledge mining, machine learning, and content personalization, with a focus on energy saving and human-centric systems. He has received more than 1200 citations. He has extensive experience in European and international research projects mainly in knowledge management, data mining, machine learning, artificial intelligence, and personalized recommendation systems, with applications in energy and transport (e.g., Teaching (H2020), Master's (H2020-Marie Skłodowska-Curie Action), (EM)3 (Qatar National Funding), and Fortissimo (FP7)).

PANAGIOTIS RADOGLOU-GRAMMATIKIS (Member, IEEE) received the Diploma (M.Eng.) and Ph.D. degrees from the Department of Informatics and Telecommunications Engineering (now Department of Electrical and Computer Engineering), Faculty of Engineering, University of Western Macedonia, Greece, in 2016 and 2023, respectively. Currently, he is the Research and Development Director of K3Y Ltd, coordinating the technical activities and strategy of K3Y in various research and development projects, including H2020 SPIDER, H2020 SANCUS, H2020 5G-INDUCE, H2020 TREEADS, H2020-MSCA Swiftv2x, TRUSTEE, INCODE, ACROSS, UP2030, and JAUNTY. He is also a Postdoctoral Researcher with the ITHACA Laboratory, University of Western Macedonia, participating in several national and European-funded research projects, such as H2020 SPEAR, H2020 SDN-microSENSE, H2020 TERMINET, H2020 EVIDENT, H2020 ELECTRON, AI4CYBER, and DYNABIC. Moreover, he is the Co-Founder of MetaMind Innovations P. C., the first spin-off of the University of Western Macedonia. He has published more than 40 research papers in international scientific journals, conferences, and book chapters, including IEEE Transactions on Industrial Informatics, IEEE Transactions of Network and Service Management, IEEE Transactions on Emerging Topics in Computing and Computer Networks (Elsevier Publishing). His main research interests include cybersecurity and mainly focus on cyber-AI, intrusion detection, and security games. He is a member of ACM and the Technical Chamber of Greece. He was included in Stanford University's list (shared by Elsevier) of the Top 2% of Scientists in the World, in 2021 and 2022. Additionally, he has been honored with five Best Paper Awards in IEEE CAMAD 2019, IEEE CSR 2021, WSCE 2022, SEEDA CECNSM 2022, and IEEE CSR 2023. He also serves as a Guest Editor for *Discover Internet of Things* (Springer Publishing) and *Electronics* (MDPI Publishing) and contributes as a reviewer for multiple scientific journals.

PANAGIOTIS SARIGIANNIDIS (Member, IEEE) received the B.Sc. and Ph.D. degrees in computer science from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2001 and 2007, respectively. He is currently the Director of the ITHACA Laboratory, an Associate Professor with the Department of Electrical and Computer Engineering, University of Western Macedonia, Kozani, Greece, and the Co-Founder of the 1st spin-off of the University of Western Macedonia: MetaMind Innovations P. C. He has published more than 270 papers in international journals, conferences, and book chapters, including IEEE Communications Surveys and Tutorials, IEEE Transactions on Communications, IEEE Internet of Things, IEEE Transactions on Broadcasting, IEEE Systems Journal, *IEEE Wireless Communications Magazine*, IEEE Open Journal of the Communications Society, IEEE/OSA Journal of Lightwave Technology, IEEE Transactions on Industrial Informatics, IEEE Access, and *Computer Networks*. He has been involved in several national, European, and international projects. He is the Project Coordinator of three H2020 Projects, namely H2020-DS-SC7-2017 (DS-07-2017), SPEAR: Secure and PrivatE smArt gRid; H2020-LC-SC3-EE-2020-1 (LC-SC3-EC-4-2020), EVIDENT: bEhaVioral Insights anD Effective eNergy policy acTions; and H2020-ICT-2020-1 (ICT-56-2020), TERMINET: nexT gEneRation sMart INterconnectEd ioT, while he coordinates the Operational Program MARS: sMart fArming with dRoneS (Competitiveness, Entrepreneurship, and Innovation) and the Erasmus+ KA2 ARRANGE-ICT: SmartROOT: Smart faRming innOvatiOn Training. He also serves as the Principal Investigator in the H2020-SU-DS-2018 (SU-DS04-2018), SDN-microSENSE: SDN-microgrid reSilient Electrical eNergy SystEm and in three Erasmus+ KA2: ARRANGE-ICT: pArtneRship foR AddressiNG mEgatrends in ICT; JAUNTY: Joint undergraduate coUrses for smart eNergy managemenT sYstems; STRONG: advanced firST RespONders traininG (Cooperation for Innovation and the Exchange of Good Practices). His research interests include telecommunication networks, the Internet of Things, and network security. He received five best paper awards. He participates on the editorial boards of various journals.

IRAKLIS VARLAMIS (Member, IEEE) received the M.Sc. degree in information systems engineering from the Institute of Science and Technology, University of Manchester, Manchester, U.K., and the Ph.D. degree from Athens University of Economics and Business, Athens, Greece. He is currently a Professor of data management with the Department of Informatics and Telematics, Harokopio University of Athens (HUA), Kallithea, Greece. He has more than 250 papers published in international journals and conferences and more than 5000 citations on his work. He holds a patent from the Greek Patent Office for a system that thematically groups web documents using content and links. He is the Scientific Coordinator of HUA in several EU (H2020, ECSEL, REC) and Qatar (QNFR) projects and national projects. His research interests include data mining and social network analytics to recommender systems for social media and real-world applications.

GEORGIOS TH. PAPADOPOULOS (Member, IEEE) received the M.Eng. and Ph.D. degrees in electrical and computer engineering from the Aristotle University of Thessaloniki, Greece, in 2005 and 2011, respectively. He is currently an Assistant Professor with the Department of Informatics and Telematics, Harokopio University of Athens, Greece. He has more than 80 publications in international academic conferences and scientific journals. He has over 20 years of experience in participating in EU-funded R&I projects in the areas of ICT, security, and robotics. His research interests include computer vision, machine/deep learning, and artificial intelligence. He is a member of the Technical Chamber of Greece.