Decoding the Multimodal Maze: A Systematic Review on the Adoption of Explainability in Multimodal Attention-based Models

Md Raisul Kibria[a,*], Sébastien Lafond[a], Janan Arslan[b]

[a]*Faculty of Science and Engineering, Information Technology, Åbo Akademi University, Turku, 20500, Finland*
[b]*Sorbonne Université, Institut du Cerveau - Paris Brain Institute, Paris, F-75013, France*

Abstract

Multimodal learning has witnessed remarkable advancements in recent years, particularly with the integration of attention-based models, leading to significant performance gains across a variety of tasks. Parallel to this progress, the demand for explainable artificial intelligence (XAI) has spurred a growing body of research aimed at interpreting the complex decision-making processes of these models. This systematic literature review analyzes research published between January 2020 and early 2024 that focuses on the explainability of multimodal models. Framed within the broader goals of XAI, we examine the literature across multiple dimensions, including model architecture, modalities involved, explanation algorithms and evaluation methodologies. Our analysis reveals that the majority of studies are concentrated on vision-language and language-only models, with attention-based techniques being the most commonly employed for explanation. However, these methods often fall short in capturing the full spectrum of interactions between modalities, a challenge further compounded by the architectural heterogeneity across domains. Importantly, we find that evaluation methods for XAI in multimodal settings are largely non-systematic, lacking consistency, robustness, and consideration for modality-specific cognitive and contextual factors. Based on these findings, we provide a comprehensive set of recommendations aimed at promoting rigorous, transparent, and standardized evaluation and reporting practices in multimodal XAI research. Our goal is to support future research in more interpretable, accountable, and responsible mulitmodal AI systems, with explainability at their core.

Keywords: Multimodal Learning, Explainable AI (XAI), Attention Models, Vision-Language Models, Explainability Evaluation, Cross-modal Explanations.

1. Introduction

Explainability in deep neural networks refers to the study of uncovering the internal mechanisms of artificial intelligence (AI) models, making their decisions more interpretable and transparent, in contrast to their traditionally black-box and opaque nature. There currently exists no universal definition regarding explainability. However, in the

*Corresponding author.
Email addresses: raisul.kibria@abo.fi (Md Raisul Kibria), sebastien.lafond@abo.fi (Sébastien Lafond), janan.arslan@icm-institute.org (Janan Arslan)

context of our work, we adopt the definition highlighted by Guidotti *et al.*, which describes explainability as an interface that functions both as a proxy for the decision-maker and as a construct comprehensible to end users [1]. In the literature, a distinction is often made between explainability and interpretability, with the latter generally referring to the cognitive process of generating meaning (e.g., a model-based feature could represent a clinical variability). However, in this paper, we use the terms interchangeably to encompass the broader concept [2]. As machine learning (ML) models have grown increasingly powerful, diverse, and accurate, there has been a parallel surge in the development of methods aimed at explaining their decisions [3], driven by growing domain-specific demands [2, 4] and increasingly stringent regulatory requirements [5, 6]. Nevertheless, explainable AI (XAI) remains an elusive goal for several reasons: the absence of a unified definition of a "valid explanation", inconsistent reporting practices, differing and often subjective stakeholder needs, and the lack of universally accepted evaluation metrics. These challenges are further amplified in the case of multimodal ML models, which simultaneously process and learn from multiple data sources. Multimodal models vary significantly in terms of the number and types of modalities, fusion mechanisms, task objectives, and application domains. Consequently, they introduce additional complexity in both the generation and evaluation of explanations. Current research on explainability for such models mainly focuses on interpreting inter-modal interactions and identifying suitable evaluation criteria. However, comprehensive studies that systematically assess and compare existing explanation methods for multimodal models are still scarce, despite their importance for standardization and benchmarking efforts.

In the last decade, advances in storage capacity, internet infrastructure, and computational resources have facilitated the creation of numerous large-scale multimodal datasets across various domains, such as COCO [7], GLUE [8], ROCO [9], VQA [10], Visual Genome [11], and MIMIC [12], among others. However, modeling these datasets presents substantial challenges, including increased task complexity and architectural limitations such as training collapse, non-convergence, and instability, particularly in generative adversarial networks (GANs) [13]. These tasks often involve processing long sequences of intricately connected tokens, which traditional models struggled to handle effectively due to limitations in contextual understanding [14, 15]. The introduction of the attention mechanism provided a breakthrough by enabling the generation of contextually weighted vectors that conditionally link input tokens to downstream modules. Originally developed for neural machine translation (NMT) [16], attention paved the way for the transformer architecture, which streamlined and generalized the attention mechanism [17]. Transformers eliminated the need for recurrent or convolutional layers and instead relied solely on attention to model relationships within and between input and output tokens. This made them highly adaptable to other modalities such as vision [18], video [19], and various multimodal tasks, thereby significantly advancing the field of multimodal learning (MML).

Modern attention-based multimodal models have achieved substantial progress across diverse application areas, demonstrating improvements in benchmark performance [20] and generative modeling capabilities [21]. In addition to their accuracy and scalability, attention-based models offer unique opportunities for explainability. Beyond traditional model-agnostic approaches, attention weights offer a model-specific lens through which token interactions can be interpreted, potentially yielding meaningful insights into model behavior [22]. However, most explanation efforts have focused narrowly on self-attention mechanisms that capture intra-modal relationships. Other critical components of trans-

Table 1: Classification and Comparison of Related Literature Reviews
Legend: ✓ = Discussion included, ✗ = Discussion not included

Reference	**Review Type**	**Key Focus Area**	**Model Architecture**	**Explainability**		
				Evaluation Dataset	**Algorithm**	**Evaluation**
Xu *et al.* [20]	Holistic Review	Transformer variants used in multimodal learning	✓	✗	✗	✗
Fantozzi *et al.* [28]	Systematic Review	Explainability of Transformer models in unimodal and multimodal contexts	✗	✗	✓	✗
Rodis *et al.* [25]	Systematic Review (methodology not fully reported)	Explainability of AI models used in multimodal learning	✗	✓	✓	✓
Our work	Systematic Review (methodology fully reported)	Explainability of attention-based models used in multimodal learning	✓	✓	✓	✓

formers, such as skip-connections, responsible for a significant share of information flow, have not been thoroughly examined [23]. Cross-attention layers, which encode inter-modal dependencies by linking queries and values from different modalities, also remain difficult to interpret due to their inherent complexity. Attempts to adapt model-agnostic explanation methods to multimodal settings have been limited and often suffer from high computational costs [24]. Existing studies also tend to focus predominantly on specific modality combinations, such as vision-language tasks [25], leaving other modality pairings underexplored. Moreover, key aspects of the explanation framework, such as the interfaces used to present explanations and the methods used to evaluate them, remain understudied. While foundational work exists on evaluation frameworks for XAI [26, 27, 1], the challenges in multimodal context have resulted in fragmented, non-systematic approaches to explanation evaluation.

Given the multitude of interconnected factors involved in explaining multimodal models, a systematic examination of the current research landscape is both timely and necessary. With the rapid advancements in state-of-the-art multimodal architectures, it is important to track the evolution, promises, and limitations of explainability within this context. Previous systematic reviews by Fantozzi *et al.* [28] and Rodis *et al.* [25] have addressed related areas. Fantozzi *et al.* focus on explanation algorithms for transformer models, while Rodis *et al.* explore explainability in multimodal AI models. In contrast, our review concentrates specifically on attention-based architectures due to their widespread adoption and performance advantages in multimodal tasks. Unlike prior studies, our approach considers explainability holistically and includes additional studies that, although not strictly multimodal, utilize multiple input streams (hereby referred to as 'multichannel') or involve generative modeling. Thus, while some overlap exists with previous work, our review protocol incorporates and extends existing taxonomies, and provides comprehensive discussions on all major components of the explanation framework [29]. Table 1 highlights key differences in focus and coverage between our review and prior work.

The main contributions of this paper are as follows:

1. We present a comprehensive catalogue of multimodal applications studied over the past four years, analyzed through the lens of explainability. Alongside systematically collecting key studies, we examine the associated domains, tasks, and the evaluation datasets — a critical factor in XAI research.
2. We adopt and extend existing taxonomies to facilitate detailed discussion across three dimensions:
    - *Attention-based architectures:* Classification of studies based on how and where different input streams are fused within the model.
    - *XAI algorithms:* Categorization of explanation algorithms, along with technical analysis and application contexts.
    - *Explainability-oriented evaluation methods:* Grouping of evaluation strategies and discussion contextualized to the XAI method and use-case.
3. We identify major challenges in achieving explainability for multimodal models and provide extensive future research recommendations based on our findings.

The remainder of this review is organized as follows: Section 2 provides a comprehensive discussion of the design and methodology for the systematic literature review (SLR), while Section 3 presents an overall summary of the findings from the selected publications included in the review. All other sections focus on the qualitative synthesis of information from the publications, analyzed from various perspectives. Section 4 and Section 5 describe the combinations of input-output modalities used and the application areas along with explainability evaluation datasets, respectively. Section 6 and Section 7 systematically analyze the model architectures and explanation algorithms employed in the reviewed studies. Section 8 elaborates on the criteria used to evaluate solutions based on explainability and contextualizes how decisions made in other sections influence these criteria. Publications introducing visualization tools for explainability as interfaces are discussed in Section 9. Finally, a critical assessment of the findings, along with future recommendations, is presented in Section 10, and the conclusion is drawn in Section 11.

2. Method

This SLR is conducted in accordance with the established guidelines proposed by Kitchenham [30], focusing on the analysis of scholarly works related to the explainability of multimodal attention-based models. To comprehensively capture the multifaceted dimensions of this research area, the study formulates the following research questions:

1. What datasets are commonly used for evaluating explainability in various multimodal tasks and domains? *(Addressed in Section 5)*
2. What types of attention-based architectures have been employed in the development of models for multimodal tasks? *(Addressed in Section 6)*
3. Which explanation methods are utilized in multimodal attention-based models, and how are these methods categorized? What is the distribution of these categories across the literature? *(Addressed in Section 7)*
4. What are the key criteria that define effective explanations, and how do evaluation methodologies vary across different application domains? *(Addressed in Section 8)*
5. In what ways do evaluation approaches account for the multimodal nature of the tasks being addressed? *(Discussed in Section 10.2)*

Table 2: Searches in databases

Database	Search
Scopus	`(TITLE-ABS-KEY (multimodal) OR TITLE-ABS-KEY ("multi-modal") OR TITLE-ABS-KEY (modality) OR TITLE-ABS-KEY (generative) OR TITLE-ABS-KEY (vision) OR TITLE-ABS-KEY (auditory) OR TITLE-ABS-KEY (audio) OR TITLE-ABS-KEY (speech) OR TITLE-ABS-KEY (text) OR TITLE-ABS-KEY (language) OR TITLE-ABS-KEY (signal) OR TITLE-ABS-KEY (image) OR TITLE-ABS-KEY (coding)) AND (TITLE-ABS-KEY (encoder) OR TITLE-ABS-KEY ("encoder-decoder") OR TITLE-ABS-KEY (decoder) OR TITLE-ABS-KEY (attention)) AND TITLE-ABS-KEY (transformer) AND ((TITLE (explain*) OR TITLE (interpret*) OR TITLE (xai)) OR ((TITLE ("tell us") OR TITLE (mean*) OR TITLE (represent*)) AND (TITLE-ABS-KEY (explain*) OR TITLE-ABS-KEY (interpret*))))`
Web of Science	`(ALL=(multimodal) OR ALL=(multi-modal) OR ALL=(modality) OR ALL=(generative) OR ALL=(vision) OR ALL=(auditory) OR ALL=(audio) OR ALL=(speech) OR ALL=(text) OR ALL=(language) OR ALL=(signal) OR ALL=(image) OR ALL=(coding) OR ALL=(code)) AND (ALL=(encoder) OR ALL=("encoder-decoder") OR ALL=(decoder) OR ALL=(attention)) AND ALL=(transformer) AND ((TI=(explain*) OR TI=(interpret*) OR TI=(XAI)) OR ((TI=("tell* us") OR TI=(mean*) OR TI=(represent*)) AND (ALL=(explain*) OR ALL=(interpret*))))`

By analyzing the various elements surrounding our primary research objective, along with related concepts and integrated insights, we propose a comprehensive guideline aimed at advancing standardized practices for incorporating explainability into multimodal models and streamlining their evaluation across tasks and domains.

2.1. Search

The search protocol was developed by the first author (MRK) and reviewed by the other authors (SL and JA). The PICO (Population, Intervention, Comparison, and Outcomes) strategy is followed for the formulation of the search string. We define the population to be attention-based multimodal models for any task from an arbitrary domain, incorporating any combination of numbers and types of modalities. The different explainability or interpretability methods are defined as the intervention. The outcome includes qualitative insights for model architecture, explanations, metrics employed for evaluation of the results, and the criteria used for determining the validity of the explanations. Based on this, we formulate the search term as presented in Table 2.

2.1.1. Search Method

The search and study selection were conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [31]. We utilized two major academic databases—*Web of Science* and *Scopus*—due to their extensive coverage, relevance to our study objectives, and complementary indexing, which helps mitigate the risk of missing relevant studies. The database-specific search terms, with slight modifications between platforms, are listed in Table 2.

2.2. Inclusion and Exclusion Criteria

We decided to constrain the search to peer-reviewed studies published within the period January 1^{st}, 2020 to January 31^{st}, 2024. The initial date is capped as the use of attention-based models in MML has grown substantially in recent times. Besides, the explainability of models has shifted from classic model-agnostic methods to more model-specific ones. The rest of the criteria for eligibility to be considered as part of the review are as follows:

1. The study either proposes a multimodal, generative, or multichannel attention-based model as the primary research objective or uses one as any part of the analysis.

2. The study discusses the explainability of the model to any extent, including qualitative or quantitative analysis.
3. Original peer-reviewed publications part of conference proceedings, journals, or workshops.
4. Article published in the English language.

The following criteria are used to exclude studies from the review:

1. Secondary studies, including systematic reviews, meta-analyses, and narrative reviews.
2. Editorials, opinion pieces, and commentary letters.
3. Books and gray literature.
4. Studies that are duplicates of other studies.
5. Studies that use ablation studies or other ways to validate their model without explicitly applying established explainability or interpretability techniques.

2.3. Study Selection Procedure

The study selection is done in the standard multi-stage approach. We first remove the duplicate studies from the final set of search results from the two databases. Next, the titles and abstracts of the studies are screened, and the eligibility criteria are applied to filter the results. Then, the full-text screening is conducted to further assess the eligibility of the studies. Finally, we perform forward and backward snowball sampling from the set of studies selected after the full-text screening.

During the title-abstract screening, we classify the studies into three categories: 'include', 'exclude', or 'ambiguous'. Any study without complete certainty is labeled as ambiguous to be more inclusive and is resolved during subsequent steps. To further improve the reliability of the review and confidently resolve the ambiguous cases, we used an additional human-guided large language model (LLM)-based screening process as described in the following section.

2.3.1. LLM-based Screening

LLMs are increasingly being utilized at various stages of the scientific reading and writing process [32, 33], including during the screening phase of SLRs [34]. Recent studies suggest that LLMs perform at a level comparable to human screeners, for example, in terms of classification accuracy when replicating the title-abstract screening of a prior SLR on *time pressure in software engineering* [35]. Based on this, we adopt a human-guided, LLM-assisted method to independently screen all studies categorized as either 'include' or 'ambiguous'.

For this task, we employ a quantized version of the LLaMA 3.1 model, trained on data with a knowledge cutoff of December 2023 [36]. The model is provided with instructions, eligibility criteria, and the title and abstract of each article. It is then prompted to follow the same three decision categories used by the authors during screening.

To ensure consistency and make the responses easy to manually process, the LLM is asked to assess the following eligibility criteria:

1. Whether the architecture used is multimodal, multichannel, or neither.
2. Which modalities are used (e.g., language, vision, video, audio, tabular, code).
3. Whether explainability is a primary research objective, a secondary one, or not explored at all.

Table 3: LLM-assisted Screening Results
Abbreviations: T/A = Title and abstract screening; Full-text = Full-text article assessment.
Legend: × = Irrelevant / Any case possible.

Authors (T/A)	LLM (T/A)	Final (Full-text)	Count	Percentage
Include/Ambiguous	Include	Include	47	72.3%
Exclude/Ambiguous	Exclude	Exclude	3	4.62%
Include/Ambiguous	Ambiguous	Include	4	6.15%
Exclude/Ambiguous	Ambiguous	Exclude	2	3.08%
×	Include	Exclude	7	10.77%
×	Exclude	Include	2	3.08%
Total			65	100%

4. Whether the explanations are based on attention mechanisms or other methods.

In addition to assessing each criterion, the LLM is instructed to extract relevant data to justify its screening decision. This simulates a chain-of-thought reasoning process, which has been shown to improve LLM performance on complex tasks [37]. The generated explanations are then manually reviewed by the first author to verify the categorization.

It is important to note that the LLM outputs are not used for direct inclusion or exclusion decisions. Instead, they are incorporated during the full-text assessment stage to provide additional input for the authors. Table 3 presents the screening results from the title–abstract phase (by both authors and LLM), as well as the final decisions based on full-text assessments.

As shown in Table 3, LLM-generated inputs are largely reliable. The LLM-assisted method provided usable screening suggestions in 90.77% of cases, with the remaining labeled as ambiguous. Of all predictions, only 13.85% were found to be incorrect after the full-text assessment. These results suggest that LLM-assisted screening is a promising approach, and we recommend further exploration and systematic adoption of such methods in the literature review process.

2.3.2. Snowball sampling

To enhance the theoretical validity and overall coverage of the study, we performed both forward and backward snowball sampling following the full-text screening phase, in line with established guidelines [38]. The sampling process was initiated from a seed study by Chefer et al. [39], selected for its strong relevance to the review and its placement early in the study period. Given the high number of citations to and from this work, we applied filtering constraints based on citation counts to manage the volume of candidate studies. The sampling was conducted on December 17^{th}, 2024, using only the Web of Science database, chosen for its more accessible filtering capabilities. Details of the sampling procedure, including inclusion counts at each step, are provided in Table 4.

Table 4: Parameters for Snowball Sampling

Sampling direction	Population size	After exclusion criteria	Percentage considered	Top cited candidates	No. of duplicates	After full-text evaluation
Forward	94	62	15%	9	2	0
Backward	50	17	30%	5	0	1

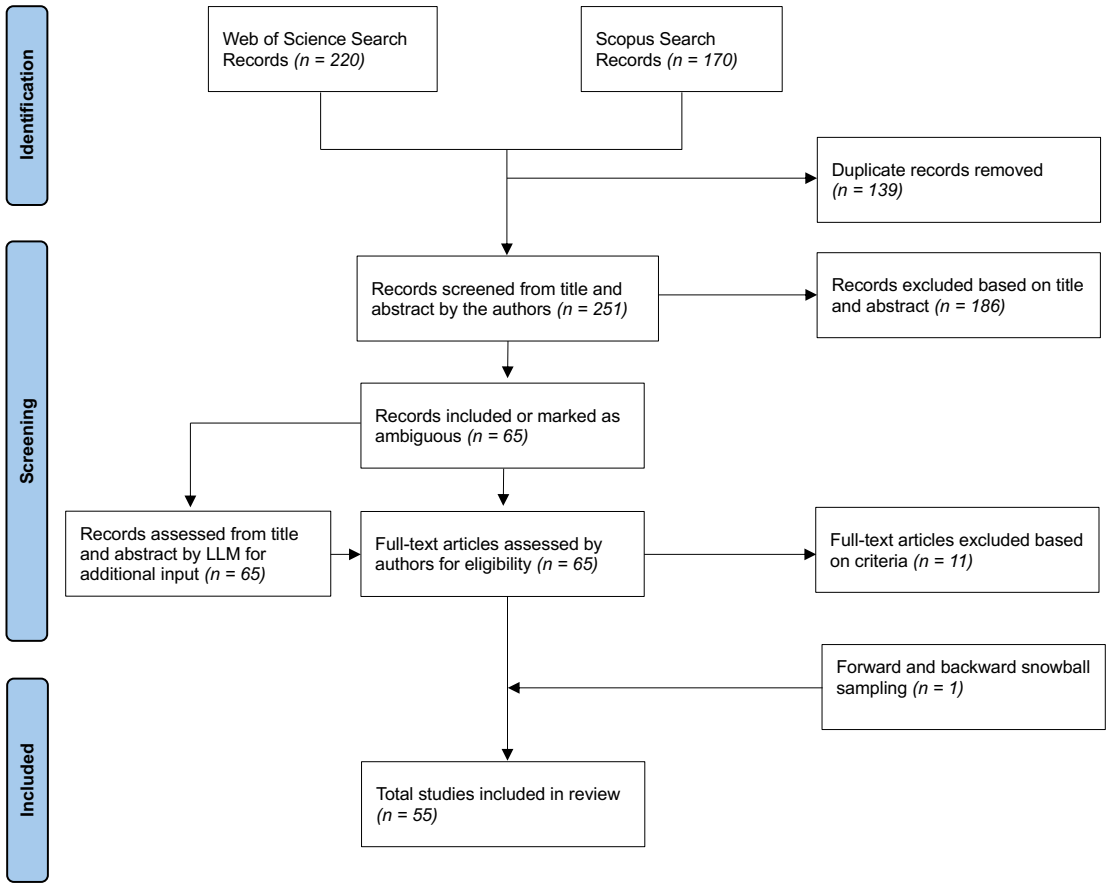


Figure 1: PRISMA flowchart for the number of selected studies during the different steps of the selection process.

3. Qualitative Synthesis Overview

After a rigorous selection process, the final approved set contained a total of 55 publications. A summary of the selection procedure is presented in Figure 1. The complete review dataset, including all computational procedures and results, is available in the supplementary material and online at: `https://decoding-the-multimodal-maze.notion.site/`. Data have been extracted from these studies for the analysis of how attention-driven multimodal models are adopted and how explainability is explored in different tasks in a diverse range of application domains. In addition, an overview of the important bibliometric analytics for the final set is presented in this section.

Primarily, the evolution of studies focused on the field over the selected period is presented in Figure 2. There has been a rapid surge of studies in the last two years, which coincides with the significant development in attention-based multimodal models over the past years and general growth in works building on previously introduced explanation methods [28]. This increased interest can also be influenced by several other factors, such as stricter regulations on AI [5, 6] and rapid mainstream adoption of AI-based agents (e.g., LLM-agents).

Additional analyses on thematic categorization, geographical distribution, and top publication venues are presented in Figure 3. For a structured thematic analysis, studies

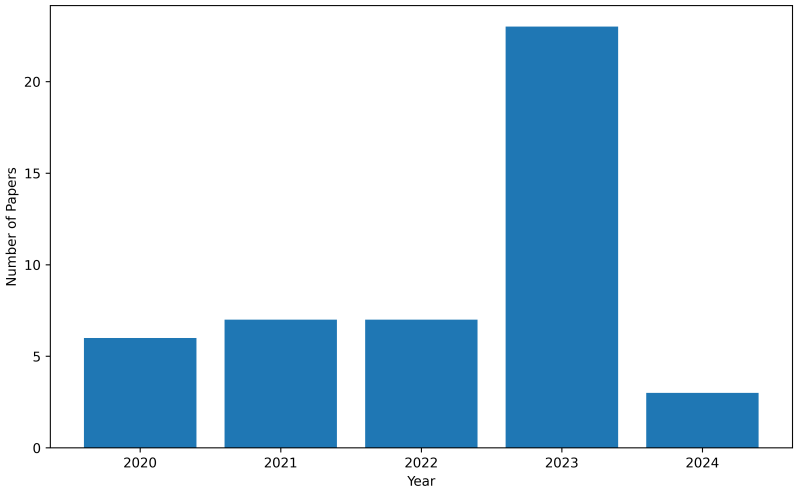


Figure 2: Publication per year

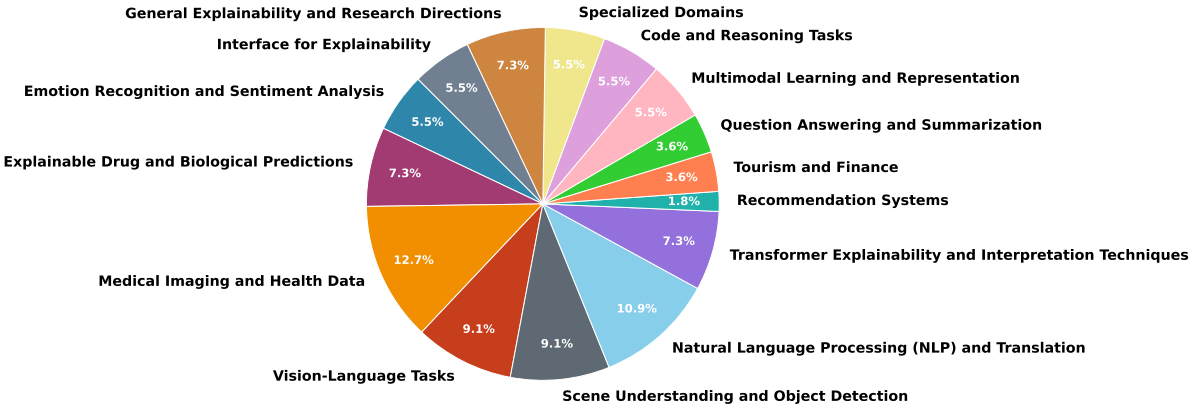
were grouped into themes defined by application domains and tasks. Although each study was assigned to a single theme to streamline the analysis, it is important to note that both the themes and assigned studies are not strictly mutually exclusive and may span multiple disciplines (e.g., *Natural Language Processing (NLP) and Translation* vs. *Question Answering and Summarization*). The thematic distribution, shown in Figure 3a, highlights the diversity of research interests across domains that demand explainability [3].

Likewise, the publication venues are varied, encompassing major conferences and journals in AI and computer vision. The most represented venue is EMNLP, the flagship conference in NLP (see Figure 3b). The geographical origin of the research is illustrated in the choropleth map in Figure 3c, revealing that the majority of publications originate from institutions based in the People's Republic of China, followed by the United States. Beyond these two countries, the distribution reflects a broad and globally diverse research landscape.

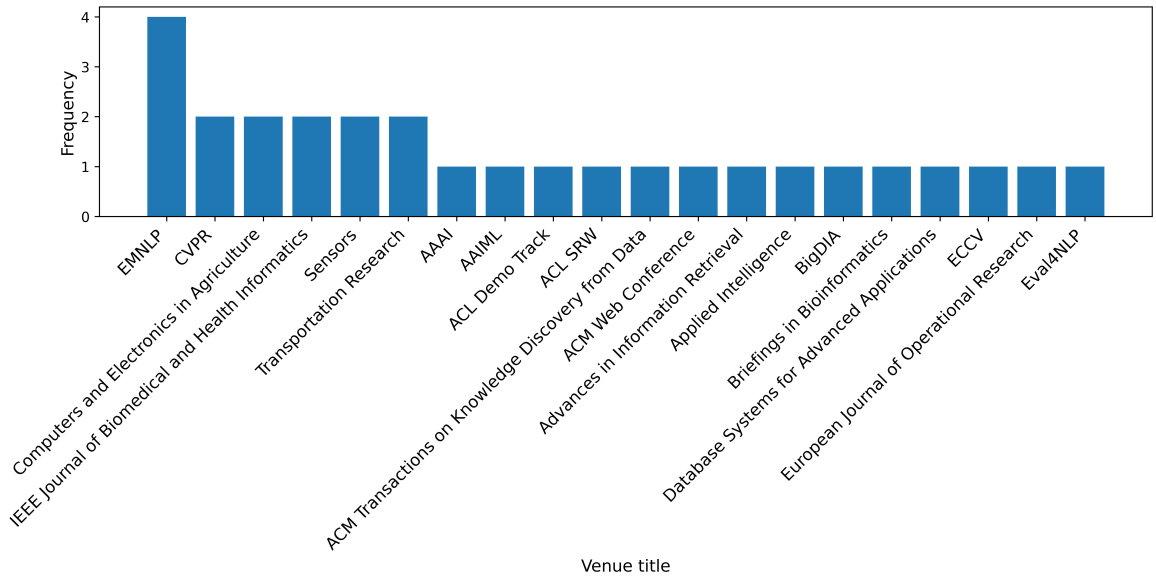
4. Modalities

Due to the wide application area considered for the generality of our study, the publications for this review reflect a diverse combination of modalities. We follow the data modality characterization by Nauta *et al.* [40] used with slight exceptions for labeling the combination of modalities. The time-series data are labeled to the same category as tabular/structured due to similar representation used for models. In addition, we make a distinct category for representing programming code, due to the difference in syntax and semantics compared to natural language. Finally, there is a separate category for audio/speech data. We adopt a flexible approach to modality labeling, recording a modality for a given publication whenever the input, output, or any intermediate representation relevant to explainability aligns with our predefined categories.

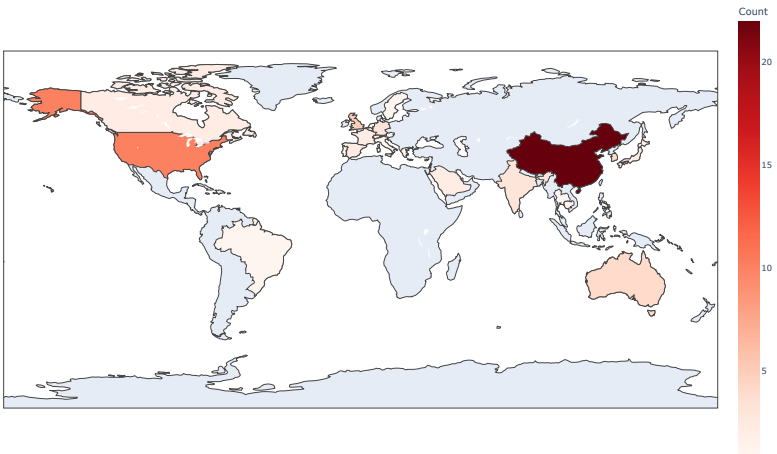
The different combinations of modalities observed in the review are shown in Figure 4b. These combinations are an aggregation of all types of data from the input and output space. The most common tasks in the area of multimodal explainability are targeted at vision-language and NLP tasks. As an individual modality of data represented either as input or output, language is quite significantly represented, with vision being the second.



(a) Thematic Distribution

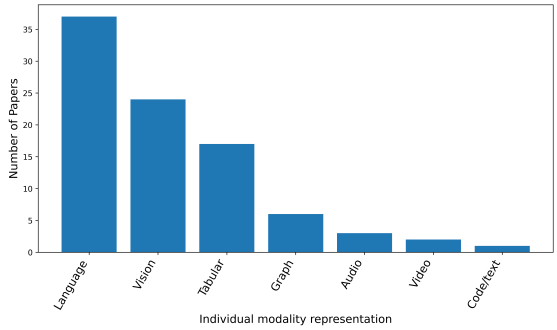


(b) Top 20 Publication Venues

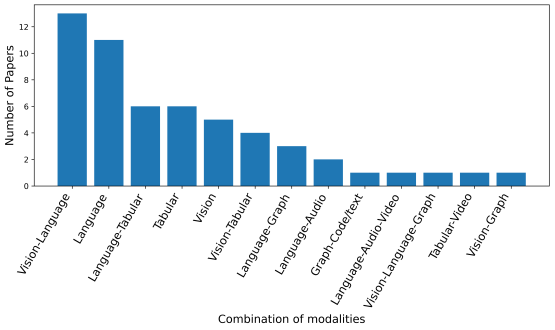


(c) Choropleth Map of the Publications

Figure 3: Key Bibliometric Analytics of the Publications



(a) Distribution of papers represented as individual modalities



(b) Distribution of papers by combination of modalities

Figure 4: Representation of modalities

The least amount of work is targeted towards code modeling. These findings are aligned with other surveys on XAI (e.g., [40]).

An important decision regarding the eligibility criteria in our study is the inclusion of "multichannel" modeling approaches. These criteria cover models that make decisions based on multiple inputs generated by processing the same source input. For instance, Yang *et al.* introduce a remote scene classification model that globally models multiscale representations from both spatial and frequency domains of an input remote scene image [41], hence making the approach multichannel and eligible for our study. The motivation behind the inclusion is that explaining multichannel approaches requires generating explanations from different representations and from different portions of the model. The other categories include multimodal models, which process and generate outputs using multiple modalities, and generative models, which produce sequences, typically within the same modality as the input. We represent these two categories by grouping them as one. The distribution of multimodal or generative and multichannel approaches is presented in Figure 5. Almost 69% of the publications in our dataset belong to the category of multimodal/generative models and 31% of them use a multichannel approach before attempting to explain model decisions.

5. Tasks and Datasets

Explainability in the set of publications is often explored for one or more primary models. These models are trained with primary task objectives that determine the datasets applicable (also determined by the domain) and how these are trained, including the architecture. The general categorization of the models used in the publications is presented in Figure 6. Studies in the area generally address more than one research objective. Among the observed task groups, the most common one is classification problems utilizing multimodality. Next to that, studies for which the primary objectives can be labeled as NLP, regression, or vision-language tasks are also quite common. All these groups incorporate a very diverse range of tasks.

The reviewed publications utilize a wide range of datasets, depending on the specific task and research domain. In general, datasets may be used for training models, validating and testing their performance, or both. Since our primary focus is on explainability, when authors specify different datasets for model performance evaluation and explainability evaluation, we report only the latter. Given the diversity of evaluation approaches (see Section 8), datasets are often not systematically used for assessing explainability.

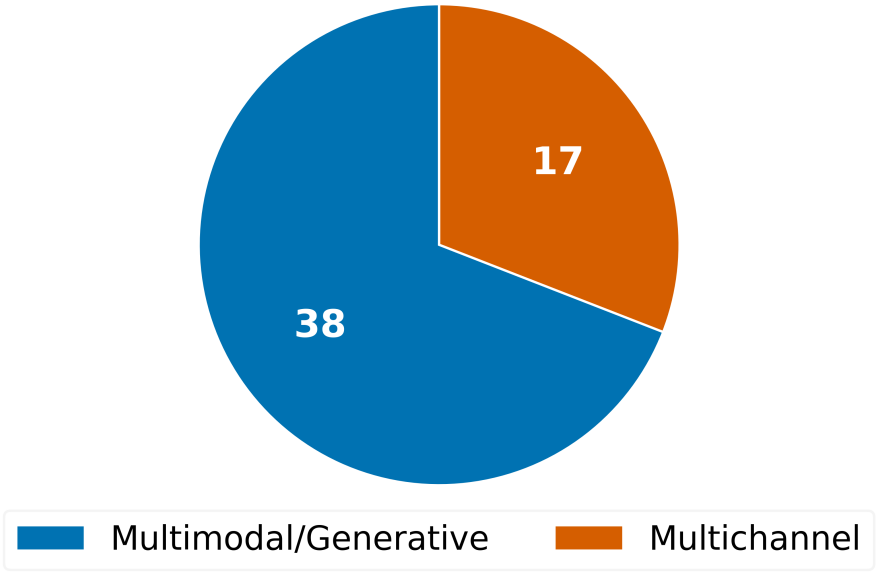


Figure 5: Distribution of multimodal/generative and multichannel modeling approaches

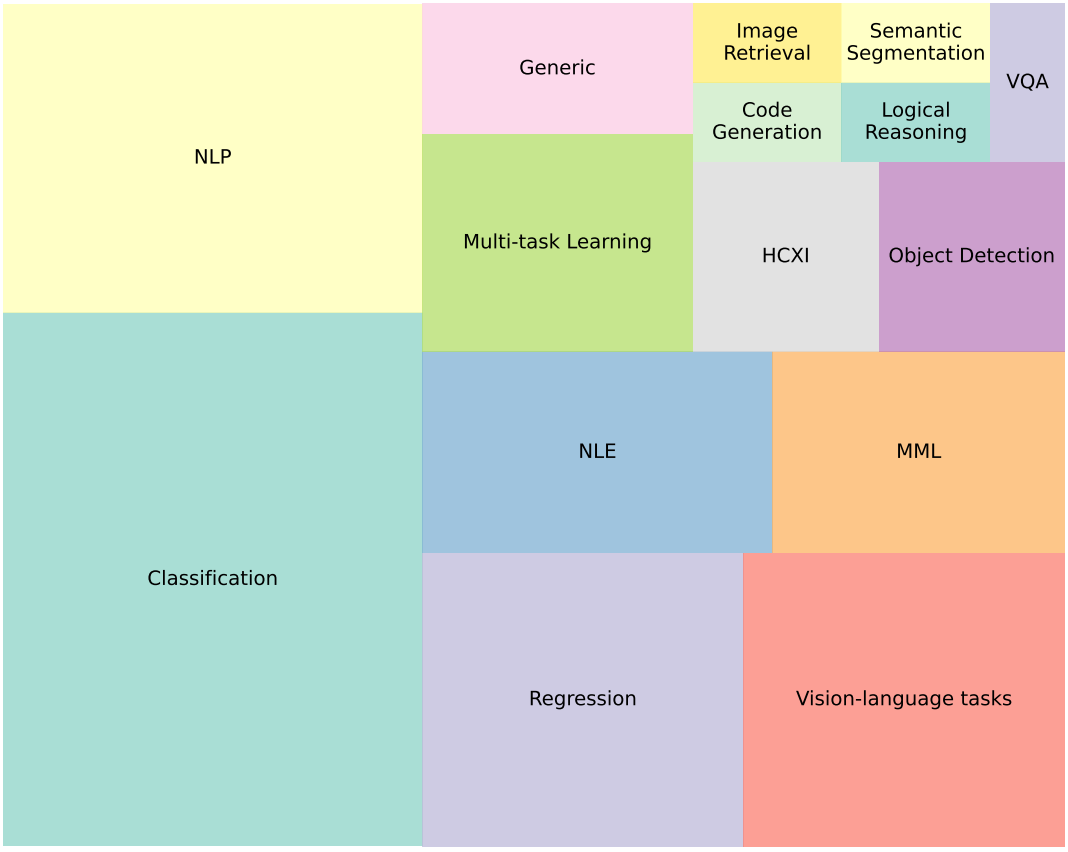


Figure 6: Primary training task objectives

Therefore, in this work, we organize our discussion of datasets based on their source tasks, with a summary provided in Table 5.

Due to the focus on explainability, a lot of studies involve tasks oriented towards explainability. NLP datasets for sentiment analysis, like Amazon polarity, SST-2, and YELP, or language inference datasets for establishing hypothesis-premise relations, like MNLI, can be leveraged for explainability-oriented studies [23]. Benchmark datasets for the visual question answering (VQA) task, such as the VQA dataset, consist of images paired with questions about their content and the corresponding answers. A later version of the dataset, titled VQAv2, included improvements to reduce language bias by including a pair of positive and negative images for each question. These datasets have also been adopted in explainability-oriented tasks in addition to being used for MML [42, 39].

Natural language explanations (NLE) refer to works on self-explaining models that can be evaluated against ground-truth labels from the datasets. For example, Guo *et al.* used sentiment analysis datasets such as YELP and Amazon-movie for personalized recommendation generation [43]. More complicated question-answering datasets, such as multiple option choice-based OpenBookQA or context-driven QA for retrieval problems-ReQA (SQuAD and Natural Questions), can be leveraged for NLE targeted tasks [44].

NLP-oriented tasks are quite diverse both in general and in the datasets used for explainability evaluation. In addition to some of the previously mention studies, annotated news corpus for abstractive summarization such as the CNN/DailyMail dataset or NYT50 by H. Wang *et al.* [45], sentiment analysis datasets like Fashion and Wine reviews by Malkiel *et al.* [46], or word alignment datasets with curated gold-standard alignment named Gold Alignment Dataset by Ferrando and Costa-jussà [47] are examples of the NLP-task group. Datasets such as the autonomous vehicle data, Lyft, or MLQE-PE for machine translation, are purely oriented for regression-driven studies [48, 49]. Datasets used for classification problems in this context often also involve solving other complex tasks such as object detection [50] or regression [51].

Large-scale datasets like MS COCO and PASCAL VOC are extended for evaluating explainability for semantic segmentation works [52, 39, 42]. These datasets are particularly valuable due to their vast sizes and the semantic annotations that can be leveraged for evaluations. VCR dataset for visual commonsense reasoning specifies a very sophisticated objective of cognition-level visual understanding. This dataset is used in two different studies [53, 54].

A more unique category of publications on human computer interfaces for explainability (HCXI) study interfaces for explainability and validate them via case studies (discussed in Section 9). For that purpose, Counterfact, a dataset generally designed for assessing whether, in LLMs, factual associations can be modified without affecting other facts, is used by Katz and Belinkov [55], and WebQA, a multi-hop, vision-language QA benchmark, is used by Aflalo *et al.* [54].

All of the listed datasets are publicly accessible either directly or through some approval process. The only private datasets are the study-specific ones manually collected by the authors and are used for various tasks such as vision-language [56, 57] and object detection with explanation [58].

6. Attention-based Architectures

Attention mechanism originally gained popularity in sequence modeling and transduction models paired with recurrent or convolutional units. It allowed selective usage of

Table 5: Evaluation Datasets used for Explainability

Dataset	Tasks	Paper
Amazon Polarity, IMDB, MNLI, QQP, SQuAD, SST2	XAI-focused	[23]
Amazon-Movie	NLE	[43]
APTV-99	Object Detection, Classification, NLE	[50]
aPY, CUB-200-2011, MNIST Even/Odd	XAI-focused, Classification	[59]
AVSD	MML	[60]
BDD-OIA	VL, NLE	[61]
CNN/DailyMail, NYT50	NLP	[45]
Code Datasets	Code generation	[62]
CounterFact	HCXI	[55]
Custom[*]	VL Object Detection, NLE Regression, VL	[56] [58] [57]
CUTE80, ICDAR-2013, ICDAR-2015, IIT5k-Words, SVT, SVTP, SynthTiger	Scene text recognition	[63]
Fashion and Wine reviews	NLP	[46]
fMRI neural activation, SICK, STS Benchmark, Toronto Book Corpus, UN-EWT	Classification, Regression	[51]
Gold Alignment Dataset	NLP, NMT	[47]
Lyft	Regression	[48]
Mboshi-French parallel corpus	NLP, NMT	[64]
MLQE-PE	Regression	[49]
MS COCO 2014	Semantic Segmentation MML	[52] [65], [39], [42]
OpenBookQA, ReQA Natural Questions, ReQA SQUAD	NLE, NLP	[44]
PASCAL VOC 2012	Semantic Segmentation	[52]
VCR	VL VL, HCXI	[53] [54]
VQA/VQAv2	XAI-focused, MML XAI-focused	[39] [42]
WebQA	VL, HCXI	[54]
YELP	NLE XAI-focused	[43] [23]

[*] Datasets with no public access available.

frames in encoder representations for sequence-to-sequence generation tasks [66, 67]. To improve the computational scalability, Vaswani *et al.* introduced purely attention-based transducers, stripping away recurrence, referred to as the Transformer [17]. The vanilla transformer model used sub-word level tokenization in an NMT use-case with several special tokens (e.g., for masking). The embeddings of the tokens were combined with positional embeddings for encoding the positional information to account for the removal of recurrence. This token-based approach, along with partial transduction, allowed the architecture to model interactions between not only the inputs but also between the input and output tokens. As a result, transformers, with slight variations, have been adapted to modeling a diverse range of modalities (e.g., vision, time-series, graph, etc.) in addition to language models. The scalability also allowed effective adaptation in multimodal contexts.

6.1. Background

There are different variants of attention used in the literature. We briefly introduce the major ones.

6.1.1. Self-attention

Self-attention is one of the most commonly used forms of attention and serves as a core component in the vanilla transformer encoder [17]. The input to self-attention is a fixed-

shape vector, Z_E, which is generated through an embedding layer in transformer models. This embedding vector is often combined with positional encoding (either predefined or learned) as part of the processing pipeline.

The embedding vector from the same source input is projected linearly onto three distinct matrices: query (Q_S), key (K_S), and value (V_S), calculated as in equation 1.

$$Q_S = Z_E W^{Q_S}, \quad K_S = Z_E W^{K_S}, \quad V_S = Z_E W^{V_S}, \tag{1}$$

Where self-attention, $SA(Q_S, K_S, V_S)$, in the vanilla transformer model can be defined as in equation 2.

$$U_S = SA(Q_S, K_S, V_S) = \textbf{softmax}\left(\frac{Q_S K_S{}^\intercal}{\sqrt{d_k}}\right) V_S. \tag{2}$$

Here, $\frac{1}{\sqrt{d_k}}$ is a scaling factor applied to stabilize the attention scores.

A key advantage of self-attention is that it allows every token in the input sequence to attend to every other token, thereby capturing long-range dependencies and global interactions within the data. Moreover, self-attention can be augmented with a masking operation before the $softmax$ step. This masked self-attention is particularly useful in scenarios such as auto-regressive decoders, where the model must prevent access to future tokens during training, ensuring that predictions are based solely on past and present context.

6.1.2. Cross-attention

Cross-attention facilitates interactions between tokens from distinct source sequences. Its formulation mirrors self-attention with key differences outlined in equation 3.

$$U_T = CA(Q_T, K_S, V_S) = \textbf{softmax}\left(\frac{Q_T K_S{}^\intercal}{\sqrt{d_k}}\right) V_S. \tag{3}$$

This mechanism allows one set of representations to focus on another, making it ideal for multimodal applications. Typically, for cross-attention in a single direction, the query (Q) originates from one sequence while keys (K) and values (V) come from another. However, without additional mechanisms, cross-attention does not provide global context from each set of representations in isolation, but only models their interactions.

6.1.3. Sparse attention

In self-attention, attention scores for each token are calculated from all other tokens in the sequence, which affects the encoding of the locality of references. Sparse self-attention, specifically, in case of structural sparsity, addresses this issue by restricting each query (Q_i) to selectively attend to only a subset ($S(i)$) of keys, as formulated in equation 4.

$$SPARSE(Q, K, V)_i = \sum_{j \in S(i)} \alpha_{ij} V_j, \tag{4}$$

with the attention weights α_{ij} defined as in equation 5.

$$\alpha_{ij} = \frac{\exp\left(\frac{Q_i \cdot K_j}{\sqrt{d_k}}\right)}{\sum_{j' \in S(i)} \exp\left(\frac{Q_i \cdot K_{j'}}{\sqrt{d_k}}\right)}. \tag{5}$$

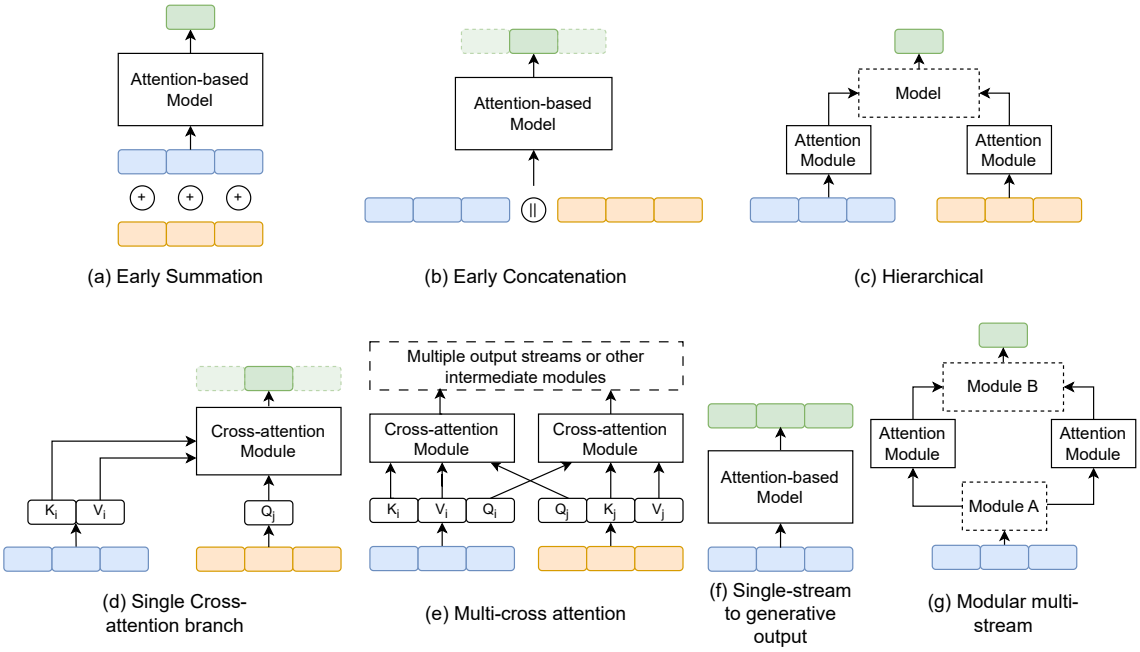


Figure 7: Block diagram illustrating various fusion architecture types: Early fusion (a, b); Hierarchical fusion (c); Cross-attention variants (d, e); and other fusion categories (f, g). Blue and orange rectangles denote input representations from different sources. Green rectangles indicate simpler outputs (e.g., classification probabilities), while a series of them suggests outputs can also be in different modalities. Shaded models or modules can have arbitrary architectures. Q, K, and V are the query, key, and value matrices in transformers.

Here, $S(i)$ is the set of key indices that the i^{th} query is allowed to attend to. The $softmax$ normalization is computed only over the indices in $S(i)$, ensuring that only the selected keys contribute to the output.

This sparsity reduces computational complexity and can be particularly useful in handling long sequences, as seen in architectures like the Sparse Transformer and BigBird. However, other variants of sparse attention for functional sparsity are also commonly used in cases of adaptive sparsity.

6.2. Multimodal Attention

Attention-based architectures, such as transformers, have been effectively extended to handle multimodal inputs and generative outputs. Several architectural variants have emerged to support multimodality, differing primarily in how and when they fuse information from multiple sources. To classify the models employed in the retrieved papers, we adapt the taxonomy proposed by Xu *et al.* [20], generalizing it from transformer-based to all attention-based architectures. This classification is particularly useful because it is agnostic to tokenization and embedding strategies, and is based instead on the fusion mechanism of input modalities. We group the original six categories into three broader architectural classes: 1) Early Fusion, 2) Hierarchical Architectures, and 3) Cross-Attention Variants. In addition to these, we observe two additional categories based on input-output structure: Single-stream to Generative Output and Modular Multi-stream Processing. The block diagram in Figure 7 illustrates the general architecture categories and Table 6 summarizes the distribution of reviewed papers across these categories.

Table 6: Architecture Variants of Multimodal Attention-based Models
Legends: * indicates studies that employ different architectures.

Multimodal Architecture	Papers	Count
Early Fusion		
Early Summation	Classification: [68].	1
Early Concatenation	Classification: [69, 70]; Regression: [71, 72, 49, 73]; VL: [65, 74, 56]; MML: [39]*, [60]; NLP: [75]; NLE: [43].	13
Hierarchical Architectures		
Hierarchical Multi-to-One	Classification: [76, 41, 77, 78, 79, 80]; Regression: [48]; NLP: [81, 82, 83].	10
Cross-Attention Variants		
Single Cross-Attention Branch	Classification: [59, 84]; VL: [53, 85, 61, 86]; Multi-task: [58, 50]; NLP: [47, 64, 87]; Code: [62]; MML: [39]*.	13
Multi-Cross Attention (Bidirectional)	Classification: [88, 89, 90]; MML: [39], [91, 42].	6
Other Architectures		
Single-Stream to Generative Output	Multi-task: [63, 51, 57]; NLP: [23, 45].	5
Modular Multi-Stream Processing	Classification: [92, 93]; NLP: [46, 44]; Segmentation: [52]; Regression: [94].	6

6.2.1. Early Fusion

A very classical method in ML for integrating information from multiple sources is through creating a joint representation before passing to the downstream model. Modern ML often uses embedding vectors as representations and are often extracted from other pretrained models. The two methods to use early fusion are as follows:

Early Summation. A simple way to fuse embedding vectors into a meaningful combined representation is to add them together. The only work using this method in the review, proposed by Meng *et al.*, is a fusion technique for their depression prediction system using the BRLTM (Bidirectional Representation Learning model with a Transformer architecture on Multimodal EHR) architecture [68]. The coded embeddings for diagnoses, procedures, medications, and topics from electronic health records (EHR) are summed together with embeddings for age and gender, and a special segment embedding for distinguishing between multiple visits. This early fusion enables access to demographic information as features along with a temporal representation.

Early Concatenation. This is a general alternative to summation when the embedding vectors come from very different sources and when semantic information between modalities is crucial. Concatenation enables the modeling of dense interactions between tokens of different modalities. This straightforward integration can be particularly useful in simpler tasks such as classification or regression. For instance, in classification tasks, an ensemble of features from pre-trained vision models can be used [70], or for medical images, relevant tabular features (e.g., encoding region volumes, cortical thickness, and radiomics properties) can be combined together with visual features [69]. For forecasting problems, contextual data can be readily provided along with the input [71, 72]. In vision-language tasks like VQA, representations of the question and visual cues can be concatenated [39, 74]. In contrast, when learning generic image-text representations, additional tokens that capture object-level information can be incorporated [56, 65]. Other works that use early concatenation are in translation quality estimation (constrained) [49], multi-granular text pair classification [20], non-covalent interaction correction [73], audio-visual scene aware dialog system [60], and recommendation generation system [43].

6.2.2. Hierarchical Architectures: Hierarchical Multi- to-One

Hierarchical architectures are designed to capture complex inter-dependencies across modalities while offering flexibility. These models employ a modular hierarchy where multiple streams representing different modalities are processed independently before being fused later in the network, often through a dedicated module. Importantly, these modules can be non-attention-based as well. With the availability of powerful pre-trained language models, hierarchical architectures are widely used for NLP and text-based classification tasks, encompassing diverse application areas. For instance, in social media it can be used for rumor detection by combining text and structured features for the context [76], in malware detection this can include combining encoding of HTTP flow and TCP streams as text and malware as image [80], and in speech emotion recognition this involves combining text transcripts with speech audio [79]. Multichannel solutions on document understanding [82, 83] and sequence generation [81] can also employ this architecture. Additionally, it supports classification tasks such as transcription factor binding site (TFBS) prediction [77] and remote sensing scene classification [41]. The only work on a regression problem by Zhang and Li is multimodal trajectory prediction by encoding environmental context images using a Swin transformer and historical context using a gated recurrent unit (GRU) [48].

6.2.3. Cross-Attention Variants

Cross-attention, by design, models cross-modal interactions. This ability allows varied implementation of the cross-attention mechanism in different architectures, such as in encoder-decoder models for the target sequence tokens to attend to the source tokens, or in multimodal models where cross-interactions within all individual modalities are required to be modeled. We further classify the cross-attention-based methods into two different classes based on how many branches of cross-attention are used in the network:

Single Cross-Attention Branch. Architectures with a single cross-attention branch only require one modality to attend to the other and not the other way around. The vanilla transformer used such an approach, which has been adopted extensively in various other tasks. The most common use of this design is in explainability-driven vision-language tasks. This includes VQA tasks where text-based justification for the correct answer is used as an additional supervision in a dual decoder setup [53], image captioning model for reason-induced autonomous driving systems [61] or captioning from visual concepts of objects [85], zero-shot image retrieval from sketches modeled as cross-modal matching problem [86]. Conceptual inputs for cross-reasoning can also be used in classification problems [59]. Another notable work by Che *et al.* implemented classification for bearing fault diagnosis, which involved training sparse attention-based multiple encoders on different scales of signal representations and combining them through a single decoder with a cross-attention mechanism [84]. In NLP, the architecture is mostly used in observing alignment in different NMT setups [47, 64, 87]. In multi-task learning, single-branch cross-attention can be used in two-stage image captioning models as a basis for recognition. The first stage involves extracting regions of interest, and the second stage generates dense captions [58] or captions with classification outputs [50]. Other uses include implementing LLMs in coding tasks (code document generation, code refinement, and code translation) [62] and MML experiments [39].

Multi-cross Attention (Bidirectional). In contrast to single cross-attention branches, multi-cross attention architectures enable interactions in more than one direction within tokens

from different source/target modalities. This design is particularly valuable for complex tasks requiring reciprocal attention and integrated modality interactions, such as MML and multimodal classification. In MML, Chefer *et al.* [65] and Bhargava [91] study the LXMERT model [95], which is a vision-language encoder for representing cross-modal learning. In LXMERT, a cross-modality encoder is stacked on top of object-relationship and language encoders for bidirectionally aligning the modalities. Bhargava replaces the $softmax$ function in self-attention with an $\alpha - entmax$ [96] for introducing functional sparsity to make the attention weights sparse. Other work using MML includes the study of the CLIPmapper model on different vision-language tasks such as image captioning and VQA [42]. For classification, application in drug discovery involves the use of structural features extracted from multiple source sequences (e.g., protein or amino acids) and processing through separate encoders [89, 90]. The architecture can also be used in multichannel classification cases such as heart failure prediction from electronic medical record data [88].

6.2.4. Other Architectures

Besides the multichannel and multimodal architectures, other key architecture groups can be defined based on how many input streams the networks handle.

Single-Stream to Generative Output. Some tasks involve generating outputs from complex spaces (e.g., image, text). The input is generally unimodal, forming a single stream of information that can be encoded into a latent representation before generating the output. Studies using such architectures in most cases are multi-task learning models, for instance, in neural encoding and decoding of distributed semantic models [51], learning physically interpretable latent representation of videos [57], or scene text recognition [63]. Other than that, this architecture can be implemented in NLP-based tasks such as controlled document summarization [45] or question-answering [23].

Modular Multi-Stream Processing. Often, an unimodal task is processed as a multimodal problem by splitting the unimodal signal into multiple streams, which are often then merged using hierarchical fusion techniques. This architecture is flexible and can be employed in various tasks. In classification, emotion recognition from selected channels of EEG [92] or from text using a dual-channel network [93]. In NLP, this type can be used for unsupervised text similarity problems [46] or evidence retrieval for QA tasks [44]. Other example uses include gene expression prediction from histopathological images [94] or semantic segmentation learned from different transformations of the same input [52].

The rest of the studies excluded from this section discuss interfaces to present explainability [97, 55, 54] and are not primarily focused on specific model-based analysis.

6.3. Discussion

Architectures play a crucial role in determining how the explainable algorithm is going to be applied, particularly when using model-specific methods. The wide range of possible combinations of modalities inherently increases the interpretive complexity, while the variety in architectures further complicates it.

In this review, the multimodal architectural groups of early summation and single-stream to generative outputs are used only by a few studies. By design, these groups only fit very specific use cases and are limited in encoding explainable multimodal interactions. On the other hand, variants of early concatenation and single cross-attention branch are more frequently used in a wide range of tasks, apart from a few exceptions, such as the

single cross-attention branch not being used for regression tasks as frequently. After classification, vision-language and NLP are two broad areas that were the most frequent in this review. Publications from the review on NLP inhibit diverse architectures having representations from almost all the architecture groups, whereas publications on vision-language tasks used either early concatenation or single cross-attention branches. Hierarchical and modular architectures are preferred for simpler tasks such as classification or regression, and are underexplored in other domains. Notably, cross-attention-based models are gaining traction, particularly in multimodal and multi-task learning. This trend may indicate a shift toward architectures that allow greater modality interaction flexibility while allowing robust interpretability techniques. Overall, the diversity in types of architecture in solving similar problems suggests that there is no specific architecture that fits well with different multimodal problems. Hence, future studies must focus on benchmarking different architecture types across tasks on cross-modal interactions and multimodal task performance with a specific focus on explainability.

7. Explanation Algorithm

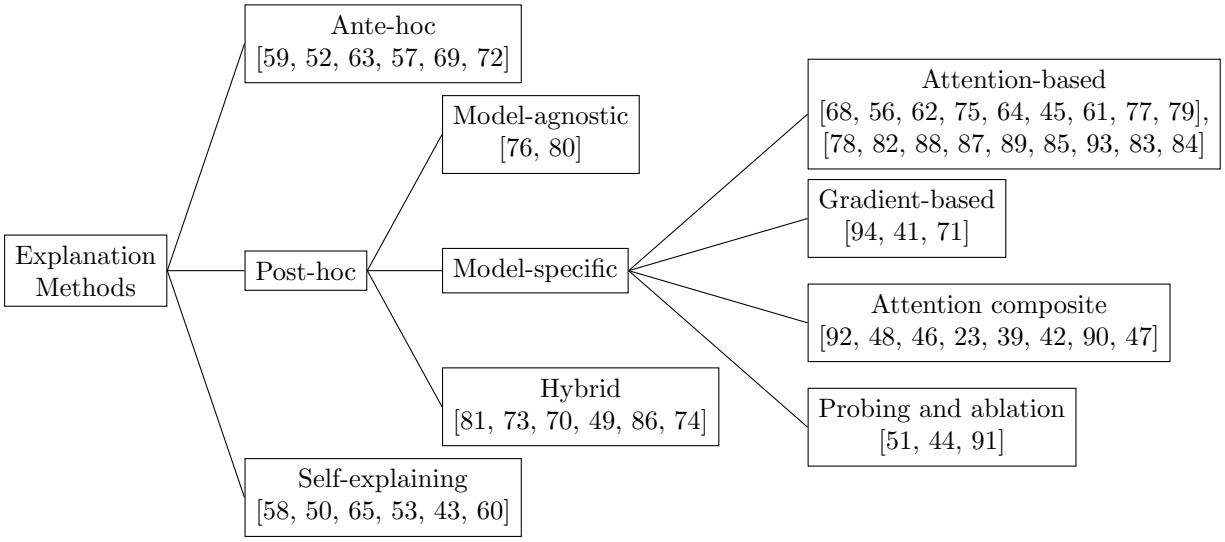
A key focus of this review is to explore how explanations are generated in multimodal applications. Due to the diverse constraints imposed by different domains and use cases, a wide range of explainable algorithms have been developed. These algorithms can be categorized in various ways, depending on the dimension under consideration [98]. In this review, we adopt a hybrid classification framework, synthesizing hierarchical structures proposed in prior studies. The primary basis for categorization is the stage at which the explanation is provided [99, 3]. The classification tree, along with representative studies, is shown in Figure 8a, while Figure 8b illustrates the distribution of papers across the classification framework.

7.1. Ante-hoc Explanations

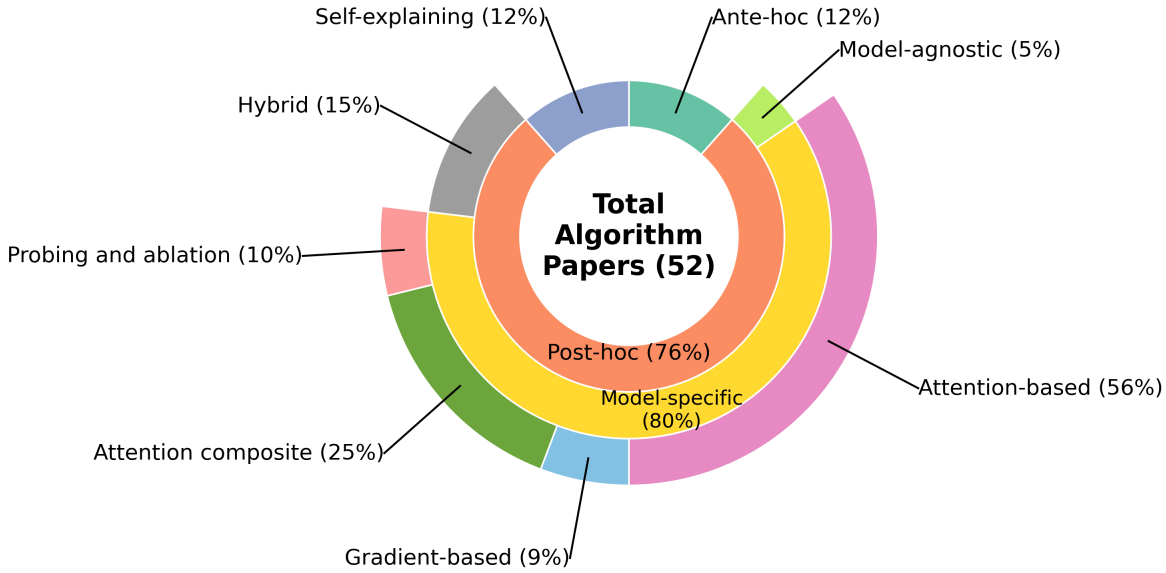
Ante-hoc explainability refers to transparency-oriented models that are intrinsically interpretable. The ante-hoc interpretability can be used for local decisions using local explainability methods. Although ante-hoc models sometimes may refer to white-box, fully transparent models, here we consider models that are designed with a focus on built-in interpretability. In this review, three publications provide ante-hoc interpretations.

Interpretability can be achieved by training models to learn high-level concepts or attributes. Concept Transformers, introduced by Rigotti *et al.*, generalizes attention weights from low-level features to externally provided high-level concept representations [59]. These concepts originate in the target domain and can be specific spatial features or global features. Concepts are learned for classification tasks without additional overhead during training. This is achieved by integrating the vector representations of the concepts directly into the cross-attention process. The query vector, Q_I, remains as the patch representation from the original input. But the key, K_C, and value, V_C, are linearly projected and concatenated representations of the concept vectors. The logits are calculated for providing the probability distribution over the output classes by multiplying an output matrix, O, with the attention outputs and then averaging over all the input patches. The formulation of the logit calculation is as in equation 6.

$$logit_i = \frac{1}{P}\sum_{p=1}^{P} CA(Q_I, K_C, V_C).O]_{pi} \tag{6}$$



(a) Classification tree of the explanation algorithms



(b) Distribution of implemented XAI algorithms for different classes

Figure 8: Classification and distribution of explanation algorithms

Yu *et al.* extend intrinsic explainability through a self-supervised learning method in eXplainable Vision Transformer (eX-ViT) [52]. The model includes two specific components—Explainable Multi-head attention (E-MHA) and attribute-guided explainer—to learn high-level interpretable attributes that are distinctly identifiable for different objects. The E-MHA calculates the attention scores in an attempt to maximize alignment with input tokens. This is achieved by introducing a scaling function (L2 norm) into the self-attention mechanism. The formulation for the proposed self-attention, SA', is as presented in equation 7.

$$\begin{aligned} A =& \mathbf{f}_\theta\left(\frac{QK^\intercal}{\sqrt{d}} + b\right)^\intercal \\ where, \mathbf{f}_\theta(x) =& \frac{x}{||x||^2} \\ SA' =& A^\intercal V \end{aligned} \tag{7}$$

The upper bound to the E-MHA ensures alignment of the attention weights, A, to the discriminative features in V, removing the distortion caused by the probability of $softmax$ as in regular self-attention. As a result, it allows a better understanding of how each token contributes to the final output. On the other hand, the attribute-guided explainer helps decompose the attention map into a set of distinct high-level attributes. This module is applied to the feature maps from the final layer of the eX-ViT's encoder for extracting trainable attributes that encode the concept of objects better. In addition, eX-ViT is trained with a composite attribute-guided loss function that combines loss values for discriminability and diversity. Discriminability loss penalizes the different types of input views, whereas diversity loss promotes effective decomposition of attribute diversity.

For scene text recognition, Buoy *et al.* introduced a single-stream to generative output architecture consisting of a ViT encoder and a 1D CTC decoder [63]. The 2D spatial feature maps (F) from the ViT are converted with a marginalization method to be used with the 1D CTC decoder. This is achieved by first passing the feature maps from the ViT through a linear layer to create unnormalized probability distribution scores with the same height and width as the feature maps, however, with the embedding dimensions remapped to the number of classes. Softmax operation is then applied to the scores across the height and the class dimensions before marginalizing over the height dimension. As a result of this marginalization method, the efficiency of CTC is preserved without losing the information of the height dimension, as is the case with 1D CTC decoders generally. The probability scores before marginalization serve as an association map (AM), which highlights the most important image regions for the prediction. Thus, along with prediction outputs, the AM provides inherent interpretability to the prediction model.

The only physics-aware neural network in this review is introduced by Kandakuri *et al.* [57] for physical representation learning from videos. The proposed model uses physical parameter extraction from frames of videos and reconstructs the frame using a spatial transformer network (STN). The differential physics layer for pose estimation ensures that the model is inherently interpretable.

Abdulkadir *et al.* used a two-stage network for medial temporal lobe atrophy (MTA) score prediction from multimodal imaging data: A ViT for accumulating image features and a TabNet classifier that processes the merged features for classification [69]. Here, the TabNet is inherently interpretable, and the feature importance from TabNet is used to

identify the most influential tabular features. Another inherently interpretable model for tourism demand forecasting was proposed by Wu *et al.* [72]. They introduce a temporal fusion transformer model, which is optimized by the adaptive differential evolution algorithm. As a result, the explainable model can be used to observe the feature importance of past and future inputs.

7.2. Post-hoc Explanations

Post-hoc explainability is designed to explain the inference process of a model after it has been trained. These types of explanations may include, for instance, providing analytics and insights, visualizing different aspects of the decision-making process, and offering explanations by example [99]. This wide set of possibilities and not requiring a change in the way modeling is done makes post-hoc explanation methods a popular choice. We discuss further classification of post-hoc explanation methods based on whether the explanation process is influenced by the choice of the model architecture.

7.2.1. Model-agnostic methods

Model-agnostic methods for interpretability are not strictly dependent on the selection of the model type and are generally applicable to any network architecture for the same task. This can be achieved by observing the inputs and outputs of the model without requiring access to the model parameters. These methods have been well-established and quite deeply analyzed in several studies [3, 98, 28]; hence, for brevity, we just mention how the techniques are used.

The most common way to achieve model-agnostic interpretability is through perturbation and by example. Local Interpretable Model-agnostic Explanation (LIME) and Shapley Additive exPlanations (SHAP) are two of the most commonly used algorithms. Janssens *et al.*, for rumor detection from social media data, explored interpretability by applying LIME due to its computational efficiency and adaptability for high-dimensional data [76]. They applied LIME on both structured and unstructured data through both interpretable and non-interpretable representations. The result of the produced explanations was local and global feature contributions. SHAP was adopted for explaining the malware-detection model by Ullah *et al.* [80]. SHAP values for the total of 32 features were used to explain how each feature influenced model decisions compared to the expected predictions. Also, SHAP was used to identify feature contributions when restricted to a specific class (benign vs. malware).

Despite their prominence in the field of XAI, implementations of these techniques in multimodal contexts remain limited. Moreover, other perturbation-based approaches are often used in conjunction with model-specific methods.

7.2.2. Model-specific methods

Model-specific or the decompositional approach generates explanations from the internal structure, parameters, and feature representation of the prediction model. As this review focuses on attention-based methods, most studies applying model-specific XAI methods leverage the attention weights. The use of attention weights as explanations, although debated [100], allows the generation of intuitive methods of observing the internal mechanics of a model [22]. Apart from attention weights, there are several other methods found in the literature for generating model-specific explanations. These methods are described following the classes used by Fantozzi *et al.* [28].

Attention-based methods. The attention scores are a matrix of cross-token probabilities. Hence, these can depict the amount of influence the tokens have on each other with respect to the final prediction. Observing the self-attention interactions was demonstrated by Vaswani *et al.* when introducing the vanilla transformer model [17]. However, a major issue with the interpretability through attention is the aggregation of the matrices from different attention heads and layers.

Final layers of neural networks are known to encode high-level, abstract features that best represent the input-output relationship. Consequently, one of the most common attention-based XAI methods is to analyze the attention scores only from the final layer. In their study, Meng *et al.* visualized the attention weights of the BERT-based encoder, which express the latent associations within EHR data of patients for depression detection [68]. The work uses the BertViz tool [101] to depict the attention component from the last layer of the network; however, the head aggregation mechanism is not specified. Ding *et al.* visualized the attention weights from the final output layers of both the sequence processing and shape processing modules. They then visualized the attention weights from the final output layers after fusing the two TFBS streams [77]. Additionally, they presented the attention weights of all eight heads in the transformer component of their DeepSTF model. Interactions between TFBSs were illustrated by varying color intensity in proportion to the attention strength. For their DLAC model, Feucht *et al.* demonstrated explainability through "top attention scores" [78]. For different target ICD-9 classes, the attention scores reflect how different parts of the discharge summary (one out of two input streams) are significant. However, the method for calculating the top attention scores is not specified. For the explainability analysis of their code models, Mohammadkhani *et al.* used normalized attention scores (in [0,100] range) by averaging the values in all the decoder layers and reporting them for each category of code tokens grouped by their types (naming, structural, others) [62]. In their work of controlled abstractive summarization, H. Wang *et al.* proposes the use of an interaction matrix, Q^s, where s denotes the sentences in the input [45]. These matrix incorporates the directional influence, q_{AB}, of sentence A on sentence B where q_{AB} is measured as follows:

$$\begin{aligned} q_{AB}(h_A, s_A, h_B, d) = \sigma\Big(& \underbrace{\boldsymbol{W}_c h_A}_{\text{informativeness}} + \underbrace{h_A^\top \boldsymbol{W}_r d}_{\text{relevance}} \\ & + \underbrace{h_A^\top \boldsymbol{W}_s h_B - h_A^\top \boldsymbol{W}_n \tanh(a_A)}_{\text{novelty}} + \boldsymbol{b}_{\text{matrix}} \Big) \end{aligned} \tag{8}$$

Here, W and b are trainable parameters, h denotes the representation of the respective sentence, and a is the summary representation accumulated with respect to the current sentence. The use of attributes specific for informativeness, relevance, and novelty gives the matrix a meaningful representation, and hence, the matrix can be visualized for analyzing each of the attributes.

Averaging the weights over all the attention heads is a commonly used method implemented in techniques such as attention rollout [22]. Dong *et al.* used averaged attention maps across all heads of the final cross-attention layer in the decoder to illustrate cross-modal interactions as a form of explanation [61]. Che *et al.* gathered the attention weights of each head from different layers and averaged them into a single matrix reflecting patch importance [84]. This helps visualize all three encoders and the decoder weights. Similarly, Y. Huang *et al.* [88], F. Xu *et al.* [82], and S. Xu *et al.* [75] used average attention scores from the last attention layer to visualize model rationale. In each

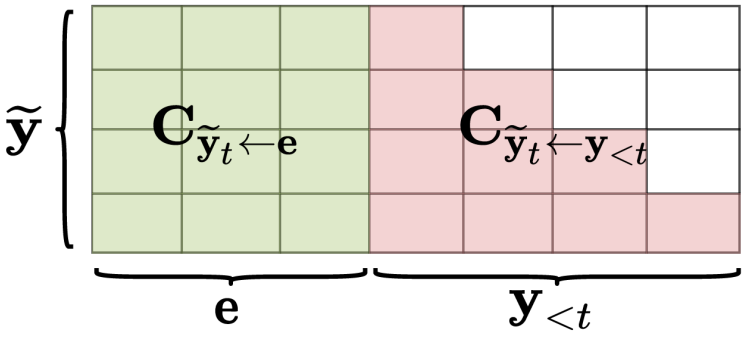


Figure 9: Decoder Layer Contribution Matrix in ALTI+ Method [87]

of these, interactions between all modeled input streams are covered. Other than these, Zheng *et al.* also presented the interpretability of their work on drug-target interaction by visualizing the attention weights [89]. Specifically, they visualize the attention scores of the input protein sequences as a 2D heatmap, but the aggregation technique is not explicitly mentioned.

The Aggregation of Layer-wise Token-to-token Interactions (ALTI) method, proposed by Ferrando *et al.* [102], calculates the token contribution matrices in transformer encoders in a way similar to attention rollout by forming a directed graph of information flow. However, instead of using raw attention weights, it quantifies contributions using the Manhattan distance between linearly transformed input and output vectors, which in turn is a measure of the information flow. This method was extended to decoder cross-attention in NMT as the ALTI+ method [87]. ALTI+ method decomposes the cross-attention output into two parts: the encoder outputs via the cross-attention mechanism and the contributions of prior decoder outputs via the residual self-attention connection. The final decoder contribution matrix, $[C_{\tilde{y}\leftarrow\mathbf{e}}; C_{\tilde{y}\leftarrow y_{<t}}]$ is then formed by combining the transformed encoder output ($\mathbf{e}$) and residual ($y_{<t}$) contributions, as presented in Figure 9. Here, each row in $\tilde{y}$ represents cross-attention outputs at different decoder time steps.

Other studies process and present attention scores differently from these methods for interpretability. Boito *et al.* used soft-alignment as the interpretation technique for low-resource NMT [64]. The attention mechanism used in this case produces a soft-alignment matrix that provides alignment interpretability between the target and source sequences. The proposed attention mechanism works as in equation 9.

$$\begin{aligned} \alpha_{t,j} &= \mathbf{softmax}(\text{align}(h_j, s_{t-1})) \\ c_t &= \text{Att}(H, s_{t-1}) = \sum_{j=1}^{|s|} \alpha_{t,j} h_j \end{aligned} \tag{9}$$

Here, the soft-alignment probabilities, $\alpha_{t,j}$, are calculated by applying a $softmax$ operation over the alignment scores between the source annotation h_j and the previous decoder state s_{t-1}. The context vector, c_t, is computed from the attention, which is the weighted sum of source annotations, h_j, and the probabilities $\alpha_{t,j}$.

Often, clustering can help with the analysis of attention scores by associating additional meaning with them. For the image captioning task, Ilinykh and Dobnik defined thematic clusters of objects from their textual labels to observe how attention links (attention scores in different heads) in the image encoder form between objects of the same

cluster in varying layer depth [85]. A cluster-based interpretability analysis for the target classes is also presented by Kumar and Raman, and Kumar *et al.* [93, 79]. However, it is important to note that in both cases, attention weights were not used for the analysis; rather, they used the last three dense layers from the modular architectures. For text emotion recognition, across the experimented datasets, the dense vectors from the three layers are used to create clusters for each of the emotion classes, which are then used for calculating inter- and intra-cluster distances [93]. Similarly, for multimodal speech emotion recognition, principal component analysis (PCA) is used for each layer before creating emotion clusters for analysis [79]. Finally, for long-form document matching, although not cluster-based, Jha *et al.* proposed the use of multi-level similarity scores for better interpretability [83]. The similarity scores are weighted across different sections and chunks of the text input and are calculated using a combination of contrastive loss and a multi-headed attention layer.

Li *et al.*, for their Oscar model, discuss the concept of full attention, where tokens from object, vision, and language representations attend to one another, and partial interaction, which restricts attention to only image and language representations [56]. These are used for the same model through the use of attention masks.

Gradient-based methods. In contrast to attention maps, which reflect where the model allocates focus during inference, gradient-based attribution methods compute the sensitivity of the output to input features, thus providing class-specific explanations by highlighting input regions that most influence a particular prediction. As most modern deep neural networks are trained via backpropagation, several gradient-based methods have been established for explainability. Grad-CAM is one of the more popular methods, introduced by Selvaraju *et al.* [103]. It was originally introduced for CNNs and involves calculating the average of gradients for each filter until the last convolutional layer, making it architecture agnostic. Among the studies in this review, Xiao *et al.* used Grad-CAM visualization for analyzing important pixels in the histopathological image inputs to their Transformer with Convolution and Graph-Node co-embedding (TCGN) vision backbone [94]. Grad-CAM visualization is generated for different classes for the same input. Due to the graph-based architecture, they also visualized the relationship between patches (nodes). For their dual-channel model for remote scene classification, Yang *et al.* proposed the use of category activation map for both the spatial and frequency channels [41]. Although the specific method is not mentioned, the category activation maps are calculated similarly to Grad-CAMs.

The only other gradient-based method used is integrated gradients (IG). IG integrates the gradients of the outputs with respect to the model inputs along a path from some baseline to the inputs. Chiewhawan and Vateekul used IG to determine the positive and negative word predictions in Thai stock market index prediction from the text part of the multimodal input [71].

Attention-centric composite methods. While attention scores offer a valuable means of modeling inter-token interactions, they often produce diffused attention maps due to the absence of localized perception. To address this limitation, complementary techniques—such as gradient-based or perturbation-based methods—can be integrated with attention mechanisms for more precise interpretability. A notable approach, called transformer attribution, in this direction was proposed by Chefer *et al.* [104], who combined LRP with Grad-CAM-style attributions derived from attention scores to enhance explanation quality. Relevance score, $R^{(k)}$, is assigned through generic Deep Taylor Decompo-

sition to each transformer block, k. These scores are then combined with the gradients, $\nabla M^{(k)}$, computed with respect to the attention scores $M^{(k)}$ through the Hadamard product. The final relevance map, aR, is produced following equation 10.

$$\begin{aligned} \tilde{M}^{(k)} &= \mathbb{E}_h(\nabla M^{(k)} \odot R^{(k)})^+ \\ aR &= \tilde{M}^{(1)} \cdot \tilde{M}^{(2)} \cdots \tilde{M}^{(K)} \end{aligned} \tag{10}$$

Here, $\mathbb{E}_h$ is the average attention relevance of multiple self-attention heads.

Transformer attribution was used by Du *et al.* to generate interpretations of their transformer-based emotion recognition model [92]. The relevance scores are calculated by taking the row-wise sum of the relevance maps, which are then also used for the channel selection task. EEG topography of the relevance scores serves as the interpretation technique. For trajectory prediction from multimodal traffic data, Zhang and Li used a modified version of transformer attribution for explaining the Swin-based image transformer [48]. The attribution method is adapted for the regression problem by replacing the predicted class labels with the predicted trajectory with different confidences. In addition, they introduce an up-sampling layer in the transformer model that can capture finer details in the attention maps, and a down-sampling layer to recover the smaller scale weight to be used during training.

Malkiel *et al.* presented BERT Interpretations (BTI), an interpretable text similarity calculation method for BERT models [46]. The method combines gradient scores with respect to the similarity score between a pair of paragraphs with corresponding activation maps for calculating per-word saliency scores. The final word-pair similarity between the paragraphs is calculated by multiplying the importance of the words in their respective paragraphs and the similarity between the pair's embedding vectors, as in equation 11.

$$\begin{aligned} Score(w_{p_1}, w_{p_2}) = Saliency(w_{p_1}). \\ Saliency(w_{p_2}).similarity(w_{p_1}, w_{p_2}) \end{aligned} \tag{11}$$

Another technique inspired by Grad-CAM, called Attentive Class Activation Tokens (AttCAT), was presented by Qiang *et al.*[23]. Yet another unimodal explanation technique, AttCAT first measures the Class Activation Tokens (CAT_i^l) for the l^{th} self-attention layer by taking the Hadamard product between the token representations (h_i^l) and their corresponding gradients, calculated with respect to the outputs for each input token, i. The final AttCAT scores are then calculated by averaging across all heads and summing over all the layers. The formulation is as presented in equation 12.

$$\text{AttCAT}_i = \sum_{j=1}^{L} \mathbb{E}_h \left(\alpha_i^j \cdot \text{CAT}_i^j \right) \tag{12}$$

Where $\text{CAT}_i^j = \nabla \mathbf{h}_i^j \odot \mathbf{h}_i^j$, and $\mathbb{E}_h(\cdot)$ denotes an averaging operation.

A more multimodal approach, introduced as an extension to the transformer attribution method by Chefer *et al.*, expands relevance-based explanations to both self-attention and cross-attention mechanisms [39]. This method calculates relevance scores separately for self-attention and cross-attention layers. An example of the explanations generated by this method, along with a comparison in the context of object detection, is shown in Figure 10.

Initially, the unimodal relevance scores are initialized as identity matrices, while bimodal (cross-modal) relevance maps are initialized with zeros. The attention map update

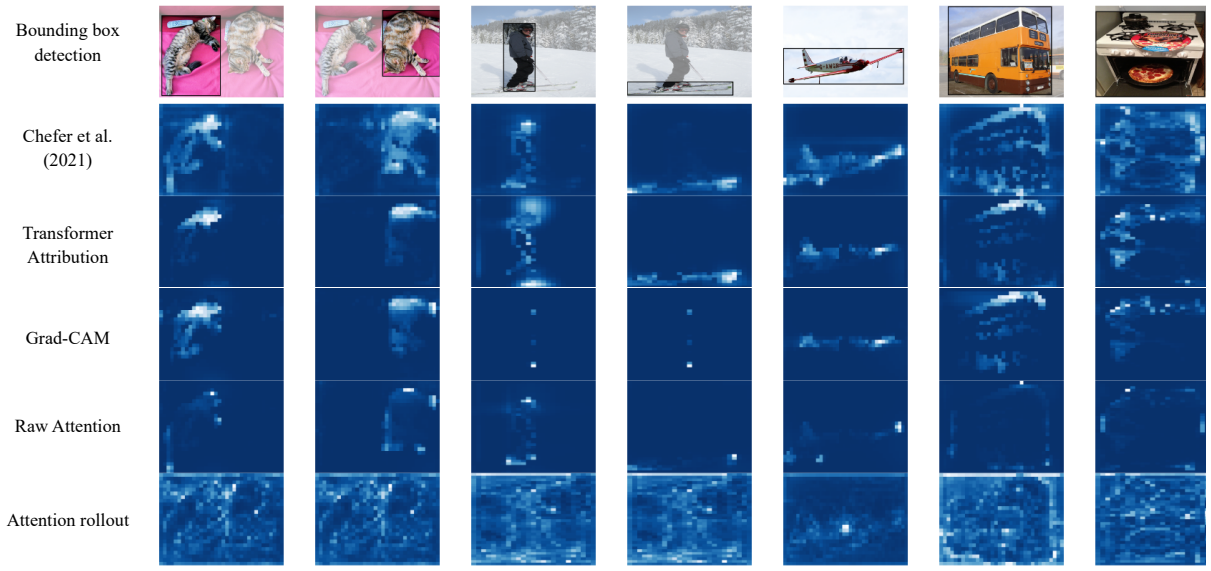


Figure 10: Visual explanation using the multimodal attention-composite method by Chefer et al. for object detection, compared with transformer attribution, Grad-CAM, raw attention, and attention rollout [39].

rule differs depending on whether the interaction occurs within self-attention or cross-attention layers. For self-attention, the relevance aggregation rule accounts for both query-query (qq) and key-query (kq) token interactions, and follows the update rule presented in equation 13.

$$\mathbf{R} \leftarrow \mathbf{R} + \bar{\mathbf{A}} \cdot \mathbf{R} \tag{13}$$

Here, $\bar{\mathbf{A}}$ denotes the attention map modified by element-wise multiplication with its gradient and clamped to positive values, averaged across attention heads to reflect the contribution of each token.

For cross-attention layers, the update rules handle relevance propagation across modalities. The key-query relevance (qk) is initialized with normalization and updated using the equations 15.

$$\mathbf{R}^{qk} \leftarrow \mathbf{R}^{qk} + \left(\bar{\mathbf{R}}^{qq}\right)^{\top} \cdot \bar{\mathbf{A}} \cdot \bar{\mathbf{R}}^{kk} \tag{14}$$

$$\mathbf{R}^{qq} \leftarrow \mathbf{R}^{qq} + \bar{\mathbf{A}} \cdot \mathbf{R}^{kq} \tag{15}$$

In this context, $\mathbf{R}^{qq}$ and $\mathbf{R}^{qk}$ represent the relevance aggregation for query-query and query-key interactions in the bi-modal case. $\bar{\mathbf{R}}^{qq}$ and $\bar{\mathbf{R}}^{kk}$ are normalized relevance matrices, and $\bar{\mathbf{A}}$ is defined as:

$$\bar{\mathbf{A}} = \mathbb{E}_h \left((\nabla \mathbf{A} \odot \mathbf{A})^{+} \right)$$

The update method of relevancy scores assumes all tokens are equally important from the outset, applying uniform accumulation without adjusting for their varying contributions. A variation of this method, proposed by Huang *et al.*, introduces an adaptive weighting strategy based on token attributions for both attention and residual connections [42].

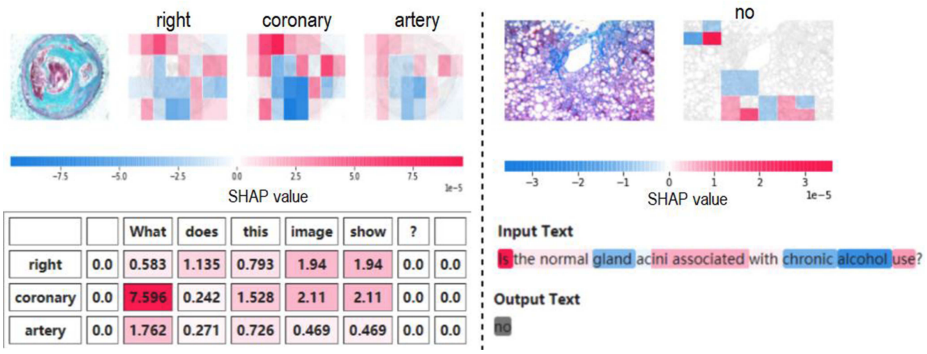
Other than these, input perturbations can be effective tools to generate explanations from the resulting changes in the attention weights. Such a method was adopted by Koyama *et al.* to interpret the interactions between T-cell receptors (TCR) and peptide–major histocompatibility complex molecules in their Cross-TCR-interpreter model [90]. They perturbed the amino acid residues in positions of interest to observe how that influenced the prediction accuracy and the attention values. In addition, they also analyzed the attention values from all four attention heads in the cross-attention layer. Similarly, Ferrando and Costa-jussà used input perturbation on the source sequence in NMT along with attention norms to generate saliency scores for each layer [47]. These saliency scores represent the contributions of source and target tokens in the prediction of future tokens.

Probing and ablation. Probing can be designed based on the task or the structure of the model to test the internal representations learned. Ablation, on the other hand, is specific to the model architecture and can incorporate the alteration or removal of specific components to understand their contribution. Although probing can also be used in a model-agnostic context, more commonly in the context of explanations, probing and ablation-based experiments are carefully designed in model-specific ways. Three studies in this review use different probing and ablation studies for interpretability. For effective visio-linguistic representations from the LXMERT, in addition to the previously mentioned adaptive sparse attention, Bhargava used Layerdrop [105] as the regularization mechanism [91]. These different elements of the model work as probes to understand how it behaves under different conditions. Firstly, a masking mechanism informs how the span of attention varies for each modality across layers. Then, changing the α parameter's value in $\alpha - entmax$ helps understand preference for sparsity. Finally, Layerdrop regularization demonstrates the trade-off between the compute runtime and the accuracy of the model. Additionally, ablation studies are conducted to conclude that adaptive span works better with denser representations of attention weights from $softmax$ in comparison to $entmax$ and that Layerdrop is effective only up to a certain depth. Overall, these provide a general understanding of how different elements of the model function across modalities and constraints, which then can be used for generating interpretations.

Similarly, Sun *et al.* explored the explainability of the encoding and decoding of their distributed semantic models through probing and ablation-based studies [51]. These examine how surface-level, syntactic, and semantic features captured by distributed semantic models contribute to modeling cortical responses during sentence processing. In contrast to the other studies, Hiemstra *et al.* proposed a lexical probing method to test the hypothesis that traditional approaches for information retrieval can outperform modern transformer-based solutions in certain cases [44]. The results from the probing experiment are used to determine methods suitable for different kinds of queries, which essentially serves as a rationale for how the resulting routing system works.

7.2.3. Hybrid

Compared to attention-centric composite methods like transformer attribution, a more decoupled combination of LRP and attention weights can be found in the work of Zanzotto *et al.* [81]. Their proposed hybrid architecture includes Kermit, a non-attention neural layer, that complements the transformer model in syntactic interpretations. To explain the encoding method of Kermit, they used an LRP-driven method called KERMITviz. KERMITviz produces heatmaps from the input syntactic tree, which provides a much better representation of causal relationships than the attention-based BertViz.



		What	does	this	image	show	?	
right	0.0	0.583	1.135	0.793	1.94	1.94	0.0	0.0
coronary	0.0	7.596	0.242	1.528	2.11	2.11	0.0	0.0
artery	0.0	1.762	0.271	0.726	0.469	0.469	0.0	0.0

Figure 11: SHAP values to explain pathology images and corresponding questions in VQA tasks [74].

A different combination of multiple explainability methods was presented by D. Wang *et al.* [73] for their multimodal model that fuses chemical properties with electron density data. The dimensionality of the extracted feature vectors is reduced using T-Distributed Stochastic Neighbor Embedding (t-SNE) for visualizing the clusters formed for different binding types. Additionally, they use plots for mean attention distances—distance between features from training samples weighted by attention—across self-attention heads, and compare heatmaps from 2D features and location maps from 3D features. The location maps are retrieved from the last block layers of their TFRegNCI and TFViT-NCI models through Grad-RAM, a visualization technique adapted from Grad-CAM for regression models. In another study, Ukwuoma *et al.* used both Grad-CAM for non-attention vision backbones and attention score visualization for their transformer encoder as an explainability method [70].

For their explainable QE task in NMT, Treviso *et al.* applied multiple explanation techniques spanning across model-specific and agnostic categories [49]. They applied attention-based (attention weights, cross-attention weights, attention $\times$ norm), gradient-based (gradient $\times$ hidden states, IG), attention-composite (gradients $\times$ attention), perturbation-based (leave-one-out) techniques as well as Relaxed-Bernoulli rationalizer. Due to the availability of a validation set for the constrained track in the shared task, these methods could be validated against the ground-truth.

Unlike Treviso *et al.*, Naseem *et al.* implemented different explanation algorithms across categories for explaining different aspects of the decision-making [74]. SHAP was used for local interpretations for both the pathology input image and the question in the pathology VQA task. SHAP provides SHAP scores of different parts of the inputs, indicating how significant these are for the model's decisions, as depicted in Figure 11. In addition, they also compared different CNN-based vision backbones using Grad-CAM visualizations for the same set of input samples. Similarly, for zero-shot matching, in addition to using self-attention maps to understand how the network learns different visual objects, Lin *et al.* used occlusion to observe reductions in similarity scores to find the most influential token pairs [86].

7.3. Self-explaining Models

An emerging category of explainable models involves systems that are trained to explain their own decision-making processes. This is often achieved through supervised training of an additional decoder that generates high-level textual explanations alongside the original task output. However, since these explanations are also produced by black-box models, their reliability remains questionable. Nevertheless, they offer a key

advantage: being high-level, they are more accessible, easier to generate post-training, and can even enhance performance on the original task [106]. One such work was introduced by Sun *et al.* with their disease detection model from images of leaves [58]. They introduce the M^2-transformer with an encoder for encoding regional information of diseases from the images and a text decoder. The M^2-transformer uses additional parameters in self-attention for improved memorization of object features and masked meshed cross-attention in the decoder for aggregating insights from different layers and different regions of the image. Two other similar works by S. Wang *et al.* and Chen *et al.* uses natural language explanations along with attention weight visualization as interpretability method [50, 65]. S. Wang *et al.* proposed a two-stage model called ODP-Transformer for pest image classification and generating corresponding textual description [50]. The first stage encodes different objects into features, which are then used to generate the captions from all the different objects and a classification output. In addition to text-based explanations, the attention weights are visualized for sample inputs to demonstrate the correlation between visual and text modalities. For image captioning, Chen *et al.* [65] propose the use of the multimodal Oscar model [56] along with attention score visualization for each token in the generated text across all attention heads.

Self-explaining models can also be useful in sophisticated application use-cases. Parelli *et al.* introduced a reasoning-guided VQA architecture [53]. The architecture consists of a BERT-based encoder for encoding the question words and a ResNet-based encoder for vision. Final attended-by-the-question representation of the image, $\mathrm{V_q}$, is achieved by taking the weighted sum of the attention weights (α^Q) and the image embedding ($\mathcal{F}$) and passing them through a linear layer as in equation 16. Attended-by-the-reasoning representation is also calculated similarly.

$$\mathrm{V_q} = \mathrm{Linear}(\alpha^Q \odot \mathcal{F}) \tag{16}$$

Reasoning supervision is achieved by aligning the attention weights conditioned on the questions with the attention weights conditioned on the reasoning. They use a two-stage training process that separates the training from the reasoning distillation, and they use attention-map visualization for explainability.

Guo *et al.* introduce a diffusion-based explainable recommendation model to generate product recommendations along with a corresponding personalized explanation [43]. Diffusion and reverse are two processes of the model used for training and inference, respectively. During the diffusion process, noise is incrementally added to the clean data of *user ID*, *item ID*, and *user comment embedding*. A transformer component learns to be robust against the noise for more expressive latent representations. Through the reverse process, the model can be randomized through a Gaussian sample to generate diverse and tailored explanations for recommended items. Lastly, Heo *et al.* used natural language for an audio-visual scene-aware dialogue generation system [60]. The multimodal architecture extracts modality-specific keywords with a GPT-2 [107] decoder for generating the explanations.

8. Evaluation Criteria

The variability in how explanations are evaluated remains a key research gap and a significant barrier to the standardization of evaluation metrics. To address this, Hedström *et al.* introduced Quantus, a toolkit designed to facilitate objective evaluation of explanations [108]. This toolkit categorizes evaluation metrics into six logical groups:

faithfulness, robustness, localization, complexity, randomization, and axiomatic metrics. In our framework, we incorporate these six categories as part of the objective evaluation class within a broader classification structure. This structure is adapted from the evaluation taxonomy proposed by Vilone and Longo [109]. We use this combined hierarchical taxonomy to guide our discussion of the evaluation metrics identified in the reviewed literature. The full categorization, along with corresponding criteria, is illustrated in Figure 12a, and the associated papers are listed in Table 7. The distribution of studies across these categories is shown in Figure 12b.

8.1. Objective Metrics

8.1.1. Faithfulness

One of the key criteria for an explanation method is how well the explanation reflects the model's decision-making. This means, for example, the important features as presented in the explanations are in fact important for the model. The quantification of the faithfulness of explanations is done in the following ways in the literature:

Fidelity/faithfulness. : A concrete example of using faithfulness for assessing interpretability was presented by Janssens *et al.* [76]. They trained a ridge regression surrogate model, g, on the interpretable representation generated by the LIME algorithm. These perturbed inputs are also used as input to the original model. Then, the coefficient of determination, R^2, is used to quantify the fidelity or how close the surrogate model's predictions are to the black-box model, hence the local interpretability. The average between the fidelity scores over N_k observations serves as the overall fidelity, as presented in the equation 17.

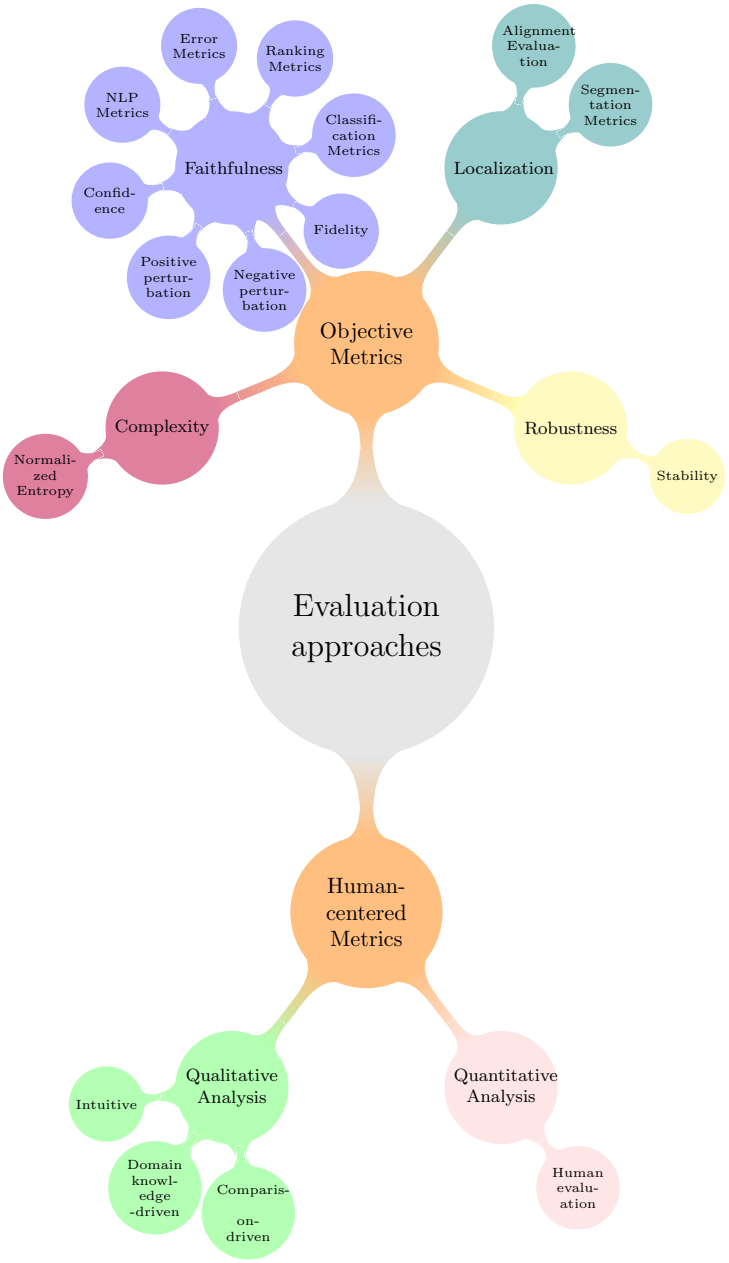
$$\text{Fidelity} = \frac{1}{K}\sum_{j=1}^{K}\left(\frac{\sum_{i=1}^{N_k} R_i^2}{N_k}\right) \tag{17}$$

Here, K represents the total number of folds of the observations.

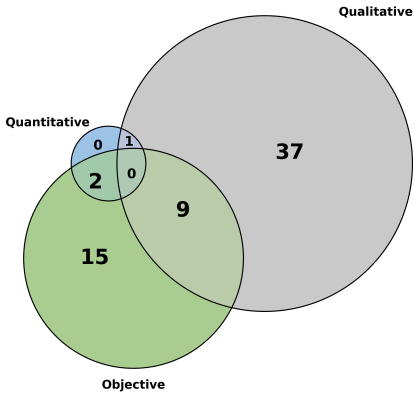
Faithfulness was also defined as one of the target explainable properties in the Concept Transformer [59]. According to equation 6, the conditional probability of each output class in the Concept Transformer is dependent on the input through the average contribution of attention scores, referred to as the positive relevance scores. Hence, these

Table 7: Overview of metric categories, criteria, and associated papers
Legends: * Indicates studies that use metrics from different criteria within the same or different categories.

Metric Category	Criterion	Papers	Count
Objective Metrics	Faithfulness	[76]*, [45]*, [39]*, [42, 49], [60]*, [87]*, [86]*, [43, 23, 59, 58, 65], [50]*, [57], [44]*, [61]*, [56]*	18
	Robustness	[76]*	1
	Localization	[47]*, [63], [39]*, [87]*, [52]	5
	Complexity	[85]*, [64]*	2
Human-centered Metrics	Qualitative	Intuitive: [41, 71, 80, 48, 62, 72, 81, 82, 75, 84, 70, 73, 79, 74], [85]*, [61]*; Domain Knowledge: [92, 89, 90, 69, 68, 77, 48], [50]*, [88, 83, 94, 74], [56]*; Comparison-driven: [93, 91, 90], [47]*, [62, 53], [46]*, [64]*, [75, 84, 51, 70, 73], [87]*, [79], [44]*, [74], [86]*	47
	Quantitative	[45]*, [46]*, [60]*	3
Non-grouped Metrics		[93, 47, 59, 85, 56]	5



(a) Categories and Criteria of Evaluation Methods



(b) Distribution of Evaluation Metrics across Categories

Figure 12: Categories of Explanation Evaluation metrics and distribution

positive relevance scores are essentially how the different ideas of concepts influence the output, making the model faithful.

Classification/ranking metrics. Classification or ranking metrics are generally metrics for evaluating the performance of models. However, when trained towards a ground-truth set of explanations, the explanations can also be evaluated using these metrics, which would then represent the faithfulness of the model. For example, in the EVAL4NLP Shared Task on QE, both in the constrained and unconstrained tracks, Treviso et al. [49] evaluated their models using Area Under the Curve (AUC), Average Precision (AP), and Recall at Top-K (R@K). These metrics were computed against word-level ground-truth labels, restricted to the subset of tokens containing translation errors. F1-score was used in addition to other metrics for evaluating the reason-induced autonomous driving system by Dong *et al.* [61]. They report the overall F1 score and mean in-class F1 score for the reasoning text generated by the model, calculated against the ground-truth. The six selected reason labels trained for result in a very limited vocabulary; hence, the reported F1 scores can reach very high values for all the tested models.

Among the ranking metrics, Hiemstra *et al.* used the mean reciprocal rank (MRR) score [110] to rank evidence for their QA task [44]. For a ranked list, the MRR metric highlights how highly ranked the relevant results are for a set of queries. The formulation is as presented in equation 18.

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \tag{18}$$

Here, Q represents the question queries.

Error metrics. When model parameters are inherently interpretable, the training loss reflects not only the model's ability to approximate the input-output relationship but also the degree to which the learned parameter values align with plausible representations of the underlying data. A significant training loss in such cases may indicate a mismatch between the model's learned structure and the inherent relationships within the data. An example of this is the work by Kandukuri *et al.* [57]. The differentiable physics layer is trained in both supervised and self-supervised fashion. The supervised loss is calculated from the error in pose estimation using equation 19.

$$L_{\text{supervised}} = \sum_{i} e(x^{gt}_{1:N,i}, x^{enc}_{1:N,i}) + \alpha\, e(x^{gt}_{1:N,i}, \hat{x}_{1:N,i}) \tag{19}$$

Here, $x^{enc}_{1:N,i}$ represents the inferred pose for each object i, which the physics engine is initiated with, and $\hat{x}_{1:N,i}$ represents the estimated pose. Both of these are compared with the ground-truth pose $x^{gt}_{1:N,i}$ using a pose estimation error function, e, calculated from the rotational error (quaternion geodesic norm) and translational error. The self-supervised loss, on the other hand, reconstructs the poses in images to calculate the reconstruction loss.

NLP Metric. Self-explaining models are trained in a supervised fashion against ground-truth natural language explanations. Hence, commonly used metrics for natural language generation tasks can be directly applied to these models to evaluate the faithfulness. The set of the metrics used include BLEU [111], ROUGE [112], CIDEr [113], SPICE [114] and METEOR [115]. BLEU and CIDEr are n-gram-based matching techniques, with BLEU

having penalization for brevity and CIDEr having additional weighing by TF-IDF values. Like BLEU, METEOR was also a metric originally proposed for measuring the quality of translations. It uses a harmonic mean of recall in addition to precision and incorporates consideration for synonyms via a thesaurus. ROUGE evaluates the recall between source and reference text, calculated based on the overlap within lexical elements. In contrast to others, SPICE is a multimodal evaluation metric used in image captioning where scene graphs are generated based on the scene and the caption, and uses a graph similarity measure to determine semantic accuracy.

In various tasks, the ROUGE metric may penalize high-quality outputs when appropriate or matching reference texts are absent or poorly aligned. Therefore, it is primarily used for text summarization-specific evaluations [45]. In contrast, many studies that utilize NLP metrics tend to rely on BLEU scores, either independently [87, 61] or in combination with other NLP metrics [50, 60, 42, 65, 58, 43]. BLEU has become a standard metric in the field of NLP due to its efficient and fast calculation capabilities, especially for comparisons at the corpus level. However, its reliance on surface-level n-gram overlaps means it often fails to assess context or fluency effectively, leading to poor performance on individual sentences [116]. As a result, in tasks like image captioning, it is commonly paired with other techniques to address these limitations.

Confidence. Confidence defines the exclusiveness in terms of the impact of explanations for each predicted class. Qiang *et al.* defined confidence as one of the desired properties of explanation [23]. The confidence of explanations per class was calculated using Kendall-τ correlation, which statistically measures the ordinal association between the saliency for classes. So, a lower Kendall correlation implies more confidence in explaining the predicted class. For a ranked array of per-token saliency, $S(x)$, the Kendall correlation for class c is calculated using equation 20.

$$\text{Kendall correlation} = \frac{1}{N}\sum_{i=1}^{N} \text{Kendall-}\tau\left(S(\mathbf{x}_i)_c, S(\mathbf{x}_i)_{C\setminus c}\right) \tag{20}$$

Here, $S(\mathbf{x}_i)_c$ represents saliency scores for class c and $S(\mathbf{x}_i)_{C\setminus c}$ is the scores for classes other than c.

This statistical measure gives a sense of how confidently explanation methods assign saliency scores for predicting each class. Although Qiang *et al.* explore confidence as a separate metric from faithfulness [23], confidence also describes the correlation between the accuracy of explanations and the predictive process. This is done by ensuring highly important tokens for one of the predicted classes have low importance for others.

Positive and Negative Perturbation. Input perturbation based on the feature importance ranks can be used to observe how the performance of the model changes accordingly. This can be done by masking features (e.g., pixels, tokens) gradually from most important to least, and observing the rate of change in performance. In case of negative perturbation, a similar process can be followed except masking features from the least important to the most important ones. Ideally, positive perturbation should result in a sharp drop in performance, which signifies that the most important features, according to the explanations, are also important for decision-making by the model. Negative perturbation should result in stable performance. Positive perturbation can reflect the sensitivity side of faithfulness, whereas negative perturbation represents the robustness side [42].

In studies by Chefer *et al.* [39] and Huang *et al.* [42], both employed positive and negative perturbation analyses for each modality in multimodal models, using the AUC as the primary evaluation metric. AUC was computed by progressively masking input tokens or image pixels along a continuous scale (from 10% to 90%) and measuring the mean top-1 accuracy at each masking level. In addition, Huang *et al.* also observed NLP metrics (BLEU, METEOR, CIDEr) for the captions in their image captioning task.

8.1.2. Robustness

Robustness of the explanations reflects how stable they are against slight changes to the model inputs, given that the output of the model remains similar. Robustness very importantly determines the consistency in explanations for similar samples, or that similar explanations result in the prediction of the same class. Expanding the metric for multimodal cases requires perturbations or counterfactual editing for all the different combinations (e.g., only one, every pair, or all) of modalities.

Stability. The term stability is often used interchangeably to represent robustness [117] to refer to the explanations holding up to slight perturbations to input instances. Janssens *et al.* describe stability as one of the most important explanation properties and adopt it for their rumor detection task [76]. They propose a new metric based on the pairwise Jaccard coefficient for calculating the global stability of feature rankings across different folds in cross-fold validation. For feature ranking sets, F_{f_1}, and F_{f_2}, from folds, f_1, and f_2, respectively, the stability across K folds and upto top-N features are calculated as in equation 21.

$$\begin{aligned} \text{Stability}_N &= \frac{\sum_{f_1<f_2}^{K} J(F_{f_1}, F_{f_2})}{\binom{K}{2}} \\ &= \frac{\sum_{f_1<f_2}^{K} \frac{|F_{f_1} \cap F_{f_2}|}{|F_{f_1} \cup F_{f_2}|}}{\frac{K!}{2!(K-2)!}} \end{aligned} \tag{21}$$

A higher value would imply more agreement on the feature importance values across different folds, resulting in more stable explanations. Stability is important to promote trust in the explanations.

8.1.3. Localization

Localization metrics assess the degree to which explanations are confined within a specified ROI. These metrics are particularly well-suited for multidimensional input spaces, such as images, and can incorporate reference ROIs defined through bounding boxes, segmentation masks, and similar methods. For models that generate two-dimensional heatmaps, localization serves as an effective measure to evaluate the extent to which saliency scores are concentrated around the original positions of target objects.

Segmentation Metrics. Tools and metrics commonly used in tasks such as object detection or semantic segmentation are inherently driven by localization. Segmentation metrics found throughout this review include mean intersection-over-union (IoU), AP, and Average Recall (AR). Mean IoU scores define how well the pixels in the heatmaps are aligned to ground-truth objects and can be used in weakly supervised semantic segmentation

studies [52]. AP and AR require the confidence scores of prediction, hence, are suitable for validating object-level heatmaps [39].

Alignment Evaluation Metrics. Alignment evaluation metric is similar to segmentation metrics and is used in both image and text localization problems. For scene text recognition, Buoy *et al.* proposed alignment evaluation by checking for overlap between the character region as indicated by the AM, R_k, and the ground-truth bounding box, Gt_k [63]. They propose the equation 22 for calculating the final evaluation score for a text with length l.

$$\begin{aligned} \mathrm{AEM}_k &= \begin{cases} 1, & \text{if } R_k \cap GT_k \neq 0 \\ 0, & \text{otherwise} \end{cases} \\ \mathrm{AEM}_{\mathrm{TEXT}} &= \frac{\sum_{k=1}^{l} \mathrm{AEM}_k}{l} \end{aligned} \tag{22}$$

For comparing the more diffused cross-attention maps to the AMs, a threshold, β, is used.

Alignments are also used in a different context in NLP, specifically in NMT. Calculated similarly, Ferrando *et al.* retrieved the alignments in bilingual NMT from cross-attention layers using their ALTI+ method [87]. These alignments were then validated against human-annotated gold alignments for calculating the alignment error rate (AER). AER helps evaluate how well the alignments are localized with respect to the reference.

8.1.4. Complexity

Complexity is the evaluation of the conciseness of explanations, which means features representative in the explanation are also the most important for prediction. A complex explanation may deem several features important, making it less accessible. This is a particularly important metric in multimodal use cases as the complexity of explanations gets aggregated.

Normalized entropy. Ilinykh and Dobnik used normalized entropy to identify the level of dispersion across the different attention heads and layers [85]. Given the source object, s_i, and the target object, t_j, in the thematic clusters and the attention distribution function, α, the entropy (E), is calculated as in equation 23. The value of entropy is measured for each attention head, h, across ℓ layers.

$$E_{\alpha_{\ell,h}}(t_j) = -\sum_{i=1}^{|S|} \alpha(s_i, t_j) \log\left(\alpha(s_i, t_j)\right) \tag{23}$$

This entropy is then normalized by dividing by the maximum possible entropy, $\log |S|$, to ensure comparability across different configurations. Higher entropy indicates a more dispersed distribution of attention weights, reflecting a higher complexity in the explanation. This suggests that the model distributes attention across multiple source-target pairs, implying a richer, more intricate relationship structure. On the other hand, lower entropy indicates saturated attention links, where a single or a few source objects dominate the attention distribution. This corresponds to a simpler, less complex explanation, as the model focuses on a narrow set of relationships.

Similarly, Boito *et al.* used average normalized entropy (ANE) to quantify the soft-alignment quality in sequence-to-sequence tasks [64]. Here, they calculated the entropy

between the source and target sentence pairs at different granularities and averaged over the tokens in the target sequence.

8.1.5. Other Quantus Categories

The remaining two categories in the Quantus taxonomy—Randomization and Axiomatic—were not utilized in any of the studies included in this review. Randomization metrics measure the rate of change in explanations as the model parameters or data labels are increasingly randomized. Axiomatic methods, on the other hand, evaluate whether the explanation satisfies certain formal properties.

Randomization is a complex method, requiring the selection of the target for perturbation, the rate of randomization, and metrics that can meaningfully capture degradation in explanation quality across heterogeneous modalities. Similarly, most existing axiomatic frameworks are designed for unimodal contexts and rely on assumptions that may not generalize to multimodal architectures. Their extension often depends on modality-specific factors, such as the fusion strategy and the nature of modalities involved. Consequently, while both categories represent theoretically important directions, their practical application in the multimodal domain remains underexplored.

8.2. Human-centered Metrics

A very effective way to determine the quality of explanations is to compare how accurately the inference process is depicted in the mental representations of the end-users. This can be systematic user studies evaluated using quantitative methods incorporating well-defined criteria. More commonly, though, non-systematic experiments along with qualitative metrics are used.

Human-centered metrics, although not as reliable or robust as the objective ones, are relatively easy to use and well-suited for capturing the subjective aspects of explanatory quality. In the literature, how these methods are implemented varies based on whether they are non-systematic and require human interpretations or systematic and incorporate user studies.

8.2.1. Qualitative Metrics

Qualitative analysis of experiments is driven by the complications involved in producing experiments for systematic evaluation and the lack of standardization in user perspectives. These often involve generic applications analyzed from intuition or for specialized tasks analyzed by domain experts. As a result, this kind of qualitative analysis is mostly done when explanations are represented at a high level, e.g., feature importance plot or heatmaps.

Intuitive analysis. High-level interpretations of general tasks, such as sentiment analysis or natural scene image classification, can be relatively accessible to lay users. For some sample inputs, the generated explanations are compared to commonly known patterns. For example, for the remote scene classification task by Yang *et al.*, the visualized category activation maps for different scenes are analyzed to observe which of the input channels approximated textures better [41]. Zanzotto *et al.* analyzed the interpretability of the Kermit model by generating explanations for carefully chosen test samples with underlying causal patterns, which the tree-like structure of Kermit encodes much better [81]. Few other studies also follow a similar method of evaluation [82].

In some cases, the underlying intuition behind model explanations is presented without adequate accompanying analysis. For instance, Wu *et al.* presented the past and

future features important for forecasting tourism demand across the three cities from the used dataset [72]. Although how the interpretations can be used is discussed, comprehensive validations of how features important for the model are relevant to the domain are not added. Similarly, for Thai stock market prediction, Chiewhawan and Vateekul presented the 10 most significant positive and negative words from the test data with no validation [71]. Similarly, Ullah *et al.* used explanations to figure out the top contributing features for malware detection [80].

In addition to these different methods, intuitive analysis can be done on the results of some experiments to generate comparative explanations. For instance, S. Xu *et al.* compare the attention matrix produced to explain the multi-granular BERT for two cases—with BERT-base by summing the attention scores for the components of the n-grams and with the same model initialized with weights from a baseline model [75]. The results of these experiments are then analyzed based on how intuitive the matrices were, their sparsity, and the attention they produced on non-relevant words. A similar method can be found in other studies [84, 70, 73, 79].

Domain knowledge-driven analysis. For more specialized domains such as medical imaging, the analysis of explanations requires deeper knowledge of the task. This might involve comparing to insights or facts established within the domain to assess the plausibility of the explanations. Such an analysis is presented by Du *et al.*, where they compare the EEG topographies with insights from neuroscience on the origins of emotion as a validation method [92]. Similarly, Zheng *et al.* [89] visually compared the 2D heatmap from protein sequences with the true binding site in the same protein. Commonly, analysis of explanations requires domain expertise in healthcare-related applications [69, 68, 88, 74] or applications in computational biology [94, 77, 90]. One of the more general applications of this method includes the document matching work by Jha *et al.* [83]. The authors selected two related documents and a non-related document and analyzed explanations for matching at different levels of detail, which is information known in advance.

Comparison-driven analysis. This group of studies typically uses case-based experiments to add credibility to the evaluation of explanations. While some papers in this area use objective metrics (as discussed in Section 8.1), they often fail to systematically measure the explanations. As a result, these approaches may rely on supplementary human-centered methods like expert reviews or intuitive analysis. For example, in their study on code models, Mohammadkhani *et al.* split the input code into different types of tokens to better understand the average attention scores across all model layers for each group [62]. They compared model predictions to gold references to classify how complex the code is, looking for patterns in specific code metrics and comparing them to trends in the dataset, such as the number of tokens or variables. Studies that use probing and ablation methods for interpretability often use this kind of analysis. In the study by Sun *et al.*, different statistical analyses and comparisons were used for determining the relative behavior of different DSM models [51]. For example, to explain the neural encoding with DSMs, they find the correlation between linguistic features and encoding accuracy for different regions of interest (ROI) from large-scale brain networks. They do further analysis on the pairwise matching accuracies, accuracy of the model in matching brain responses to the corresponding stimuli, for the same scenario with different ROI topics. Similarly, to understand how adaptive mechanisms reflect on the learning of multimodal representations, Bhargava evaluated different configurations of their sparse attention-based method against the baseline through test set (test-dev and test-std) accuracy [91].

These methods observe model behaviors and establish the constraints under which these emerge.

Such techniques are also used in novel studies revolved around explainability, such as identifying the implications of training self-explaining models. For instance, Parelli *et al.* quantitatively compared the test set accuracy of their VQA models before and after aligning to the reasoning [53]. In addition, they also used ablation accuracies to observe the drop in accuracy after masking key visual features in both cases. In their image retrieval work, Lin *et al.* used experiments like finding key matching-feature pairs by observing the maximum drop in relation score in individual matches [86].

8.2.2. Quantitative metrics - Human evaluations/user studies

Rooted in qualitative judgments, quantitative human evaluations address both the subjectivity of explanations and provide a systematic framework for comparing different models based on that. These incorporate the two categories of evaluation approaches proposed by Doshi-Velez and Kim—application-guided and human-guided evaluation [26]. With certain bias, such studies can be used to compare explanations across different modalities, various forms of explanations (e.g., high-level vs. low-level), and across multiple criteria [29]. However, one very clear issue with these is the lack of a standard protocol followed for such studies [29, 98]. Although a commonly used form of experiment, in this review, only three publications report the use of human evaluations.

For their abstractive summarization work, H. Wang *et al.* proposed human evaluation for two different tasks [45]. The subjects are lay users from the Amazon Mechanical Turk platform. In one experiment, question answering evaluations were designed so that each user answers questions devised from the gold summary while having access to the model-generated summary only. The responses are evaluated on a 3-level scale (correct, partially correct, incorrect). The other evaluation performed was to assess the generation process on four criteria: informativeness, novelty, relevance, and fluency. Users evaluate and rank summaries generated by different models for each category, ordering them from best to worst after reviewing the original document. Only the selection of the best cases or the worst cases is considered when calculating the final scores for each criterion-candidate model pair. Paired-T tests are then conducted to compare to the baseline. Although the human evaluation experiments in this study are mainly focused on the performance rather than the explainability part of the solution, for evaluating self-explaining models, these methods can be adopted.

A more explainability-oriented human evaluation was conducted by Malkiel *et al.* for their work on text similarity matching [46]. To compare the proposed solution to baseline explainability methods, they used a 5-point mean opinion score from novice users. The users were presented with randomly shuffled explanations of test set samples from the candidate methods. The users evaluate the same number of samples from each candidate from 1 (poor) to 5 (excellent), and the mean and the standard deviation of the experiment serve as the interpretation score. Similarly, a 5-point Likert scale is used for evaluation in the DSTC10 challenge [118], where the study by Heo *et al.* [60] on multimodal scene-aware dialogue systems was one of the submissions. In this challenge, five humans evaluate each answer based on their own view of the correctness while also considering the relevance, fluency, and informativeness of the answer. The final score is the average of each criterion. This experiment is again suitable for high-level explanations, such as natural language or natural images.

8.3. Other Metrics

In addition to the two established evaluation methods with standard categories, some implementations for assessing explanations do not fit within these frameworks. For instance, plausibility is a well-established metric in XAI literature, used to evaluate how well explanations align with human understanding of a task, irrespective of their faithfulness [117]. By nature, plausibility necessitates human-centered studies, as it depends on subjective judgments grounded in domain knowledge. However, Rigotti *et al.* addressed this by incorporating plausibility into their Concept Transformer as an integral part of the model's design, alongside faithfulness [59]. This was achieved by integrating human-interpretable concepts during training. Formally, given an expected attention distribution H (aligned with domain knowledge) and attention weights A, they introduced a specialized explanation loss term $\mathcal{L}_{\text{expln}}$ in addition to the classification loss. This loss is defined as in equation 24.

$$\mathcal{L}_{\text{expln}} = ||A - H||_F^2 \tag{24}$$

Here, $||\cdot||_F$ denotes the Frobenius norm. By minimizing this loss, the model adapts its weights to incorporate prior domain knowledge, thereby learning plausible, interpretable concepts.

Beyond these metrics, other approaches employing comparison-driven analyses use metrics that are not traditionally tied to explainability. For instance, studies often generate clusters of internal feature representations to establish a foundation for explanation, analyzing them via class-based grouping [93] or thematic tags [85, 56]. These analyses leverage related metrics such as average cosine similarity [47, 85], inter- and intra-cluster distances [93], or visual inspection [56].

Notably, Feucht *et al.* introduced an explanation method for ICD-9 label classification, but it is the only study in the review to not include any specific evaluation [78].

8.4. Discussion

In summary, a total of 81 instances of evaluation methods were identified across the 52 studies included in the review. Five of these instances could not be classified within the proposed classification framework due to non-standard implementation or limited suitability as formal evaluation metrics. Nonetheless, they may still offer comparative insights, particularly when analyzing internal model components such as latent vectors. As shown in Figure 12b, the majority of metric instances are qualitative. This prevalence is likely due to the relative ease of implementation, even though such methods often yield subjective or partial insights. Authors may prefer these methods because they allow for intuitive explanations of a model's decision-making process without the need for rigorous experimental design. Human-centered quantitative studies are rare in the context of multimodality, with only the three instances identified. This scarcity may be attributed to the high cost of conducting such studies comprehensively and the lack of standardized study design guidelines. Among all the evaluation instances, 40.6% employed objective metrics, and 42.3% were found in studies that also used either qualitative or quantitative methods. Despite their presence, the use of objective metrics is highly concentrated with 15 out of the 26 instances relying solely on "faithfulness" as the metric. Moreover, none of the identified objective methods explicitly quantify inter-modal interactions or assess multimodality in a holistic manner. This indicates a narrow focus within the category and highlights the need for greater diversity in evaluation metrics along with dedicated methods for future multimodal research.

9. Explanation Interfaces

Another critical aspect of explainability is its presentation. Explanations often lack the cognitive engagement necessary to foster trust and reliability. Consequently, an emerging area of research focuses on the intersection of human-computer interaction (HCI) and XAI to make AI decision-making processes more accessible [27, 29]. In this review, beyond studies that explore explainability as a primary or secondary objective, we highlight three works that specifically address the presentation and interface of explanations [97, 55, 54].

Sarti *et al.* introduce Inseq, a Python toolkit that adapts various perturbation-, gradient-, and model-internal attribution methods for both encoder–decoder and decoder-only sequence generation models [97]. In addition to a lightweight visualization framework, the toolkit supports different aggregation strategies for token-level attributions (e.g., subword-, neuron-, or layer-based grouping). The interface is demonstrated through multiple case studies. One application involves detecting gender bias in NMT by analyzing the use of occupations and pronouns to uncover spurious correlations across gendered translations. Qualitative validation is also conducted using attribution scores for sample sentence pairs. Another case study focuses on locating factual knowledge in GPT-2 using contrastive attribution tracing, which identifies layer-wise and token-wise contributions to factual content. The toolkit is well-suited for generating explanations in a range of generation tasks, including NMT, code synthesis, and dialogue generation.

In another study, Katz and Belinkov present VISIT, a tool designed to visualize semantic patterns in the information flow of transformer-based, GPT-style models [55]. VISIT projects attention weights and hidden states into the vocabulary space, rendering them as forward-flow graphs that reflect attention head activations and dynamic memory states during prompt completion in GPT-2. The tool also enables probing of other model components to better understand decision-making processes. Human-centered workflows are supported in two key ways: first, through intuition-driven analysis of an indirect object identification (IOI) task, which highlights the role of specific transformer components (as seen in Figure 13); and second, through comparison-driven analysis to examine the influence of layer normalization and regularization across different model configurations.

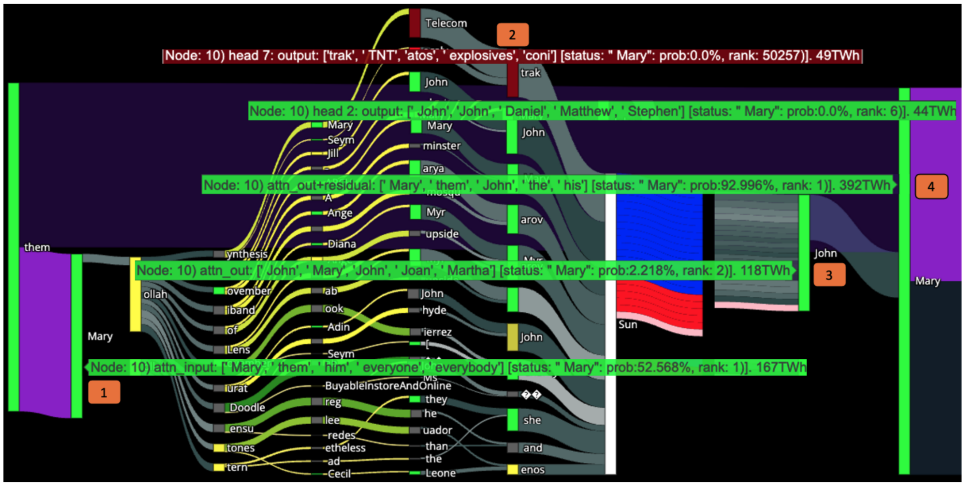


Figure 13: Flow-graph for layer 10 attention from GPT-2 small in VISIT for an IOI task [55].

Finally, Aflalo *et al.* propose VL-InterpreT, an interface for interpreting vision-

language multimodal transformer models [54]. VL-InterpreT provides visualization of both cross-modal and intra-modal attention. It comprises two main components: attention maps across all heads and layers, and hidden state plots that track the evolution of token embeddings. The interface is validated through case studies involving the KD-VLP model [119], analyzed on two benchmarks—VCR and WebQA—using attention-based metrics, cluster analysis, and heatmaps. The analysis covers both successful and failure cases, offering a bottom-up, visually guided approach to model interpretation for end users.

Taken together, these interface-focused tools transform raw model internals into intuitive and interactive experiences. They enable users to explore attention flow, detect biases, trace factual retrieval, and compare architectural variants—effectively bridging the gap between complex transformer operations and meaningful user insights.

10. Recommendation and Future Directions

Based on our analysis of the current state of explainability in multimodal attention-based models, we identify several key areas that warrant further attention. The following recommendations are framed to guide future research directions that align with these findings.

10.1. Streamlining Multimodal Architecture

As deep neural network architectures continue to evolve, so does the complexity of their design and training, particularly in multimodal settings. This trend is evident even within the narrower scope of attention-based models reviewed here. The wide variety of architectural configurations—especially the fusion mechanisms employed for combining modalities—poses significant challenges for explainability. Often, architectural choices are influenced by baseline models, leading to clusters of similar designs that are typically evaluated within isolated groups. As a result, the same task in different domains may yield vastly different model structures.

To promote generalizable explainability, it is imperative to streamline architectural choices across tasks and domains. This includes comparing multimodal models across diverse fusion strategies, such as early, hierarchical, cross-attention-based and modular fusion methods, as outlined in this review. Doing so ensures that all relevant token-level interactions are captured and provides an empirical basis for identifying the most appropriate fusion strategy in different contexts.

10.2. XAI Algorithms for Modeling Multimodal Interactions

Despite substantial progress in the development of XAI techniques, most existing algorithms remain limited in their ability to capture the complex interactions that characterize multimodal models. This is particularly evident in the case of attention-based methods—by far the most common explanation strategy observed in this review (34.6% of the total algorithm papers)—whose validity as explanation tools remains contested [100, 120, 121].

Multimodal interactions typically occur at three levels: (1) between tokens within individual modalities, (2) between token pairs from different modalities, and (3) between grouped interactions involving multiple tokens and modalities. Current attention-based and composite methods can only effectively capture the first two types, while model-agnostic methods tend to model only intra-modality dynamics.

Emerging approaches such as self-explanations, particularly those enabled by LLMs, offer high-level natural language justifications. However, as these are also generated from opaque, black-box systems, they lack the transparency and reliability in tracing model-internal decision paths. Therefore, future XAI algorithms (like the recently introduced InterSHAP [122]) must be capable of modeling full-spectrum multimodal interactions, while maintaining computational efficiency and transparency in decision-making.

10.3. Cognition-Aware and Domain-Aware Modality Fusion

Existing multimodal XAI algorithms often assume equal contribution from all input modalities, an assumption that contradicts well-established principles of human cognition. Research in neuroscience and psychology has shown that different sensory modalities contribute unequally to human decision-making depending on context and individual differences [123]. Furthermore, the criteria for interpretability are known to vary significantly across application domains [117].

Future work should focus on developing cognition- and domain-aware fusion strategies that account for such variability. Weighted fusion approaches, such as that proposed by Huang *et al.* [42], offer a promising path by enabling conditional contributions from modalities. These methods can be extended to incorporate a broader range of cognitive modalities, guided by insights from interdisciplinary research in human perception and cognitive science. To address domain variability, Islam *et al.* introduced a model-agnostic, weighted framework for quantifying explanations that can be applied across different domains [124]. Although promising, both these types of techniques are currently very limited to specific use-cases. Significant research should be carried out in this area for cognition-, task- and domain-aware modality fusion for generating meaningful and context-sensitive explanations.

10.4. Explainability as a Core Design Objective

Despite advancements in both algorithmic and evaluation methods, explainability is still often treated as an afterthought rather than a fundamental design criterion. Foundational works have emphasized the value of model-agnostic frameworks [117] and application-grounded evaluations [26] for achieving robust explainability. Yet, adoption of these recommendations remains sparse, particularly in multimodal contexts.

Algorithmically, many studies limit their scope to partial explanations, either covering a subset of modalities or analyzing each modality in isolation, typically using attention-guided, model-specific techniques. Only a few approaches address inter-modality interactions explicitly [42, 39]. From an evaluation standpoint, most works rely on ad hoc qualitative analysis, with only a minority incorporating both human-centered and objective assessments (only 17.19% of the total in this review, as seen in Figure 12b).

We emphasize that any system labeled as "explainable" must undergo extensive experimentation for both algorithmic and evaluation aspects of explainability, at the very least. As Doshi-Velez and Kim argue, multiple dimensions of explainability—such as target audience, formulation of interpretability, evaluation level, task-related factors, and XAI interface—should be clearly documented [26]. Ideally, a standardized reporting guideline specific to explainability should be developed, drawing inspiration from frameworks such as CONSORT-AI [125], to promote reliable and reproducible adoption of XAI practices.

10.5. Towards Deeper and More Systematic Evaluation

Evaluating explanations remains a challenging task due to the inherently subjective nature of interpretability and the diversity of applications and domains in which multimodal models are deployed. While several comprehensive evaluation frameworks exist [40, 117, 108, 26], very few studies apply more than one objective criterion and none address multimodal interaction metrics specifically. For example, the Quantus toolkit organizes existing objective evaluation methods into six logical categories to facilitate reproducible and accessible assessments [108]. However, in this review, only four of those categories were used, and evaluation methods outside the "faithfulness" class were rarely applied (only 8 out of 52 studies, or approximately 15.38%).

We encourage the integration of methodologies from adjacent disciplines such as cognitive science [126], HCI [127, 128], and statistics [129] to support deeper, more grounded evaluation of multimodal models. Particularly, there is a need for metrics that quantify the effectiveness of explanations in capturing cross-modal dependencies. Recent proposals in this direction include modality heterogeneity [130], multimodal matching scores [131], synergistic faithfulness and unified stability [132], though their adoption remains limited.

Additionally, human-centered evaluations remain largely unsystematic. Most rely on informal qualitative feedback, while those that involve quantitative studies often lack adherence to established protocols. Given the crucial role of human users in the interpretability loop, we echo Vilone and Longo [109] in stressing the importance of standardized, domain-sensitive human evaluations that account for application context and modality-specific cognitive load.

11. Conclusion

In this work, we have provided a comprehensive overview of explainability in multimodal attention-based models. This overview was motivated by the growing gap between explainability and the rapid advancements in modern AI applications. First, we systematically collected relevant publications in the area and analyzed them based on application tasks, domains, and evaluation datasets. We then identified the architectural variants used in the multimodal literature, with a particular focus on the fusion methods employed for combining input modalities. Based on this, we classified the models into four fusion categories and outlined the prevailing trends and limitations in architectural choices. Subsequently, we conducted an in-depth discussion of the selected publications along two key dimensions of explainability: explanation algorithms and evaluation methods. We adopted and extended a taxonomy grounded in prior research to categorize studies and highlight their distribution across these dimensions. Additionally, we briefly reviewed studies that proposed interfaces for explainability. Finally, we offered detailed recommendations and outlined future research directions based on our findings.

Our analysis reveals that, while there has been a significant body of work on multimodal explainability, the adoption of tools and methods—many of which stem from an already incomplete field of explainability—still requires considerable refinement for effective use in multimodal scenarios. This observation holds true across all examined dimensions: application-architecture, explanation algorithm, and evaluation strategy. The widespread but inconsistent use of the term "explainable solution" has further hindered progress toward standardization in the field.

As emphasized in this review, explainability must be rigorously developed, experimentally validated, and transparently reported in order to create truly explainable solutions. This is especially critical in the multimodal context, where the complexity of modeling nuanced interactions among multiple data modalities poses additional challenges. Ultimately, explainability should be a central consideration in multimodal model research, particularly as such models grow increasingly powerful. We hope that the responsible and rigorous integration of explainability—guided by the insights presented in this work will contribute to more transparent, trustworthy, and reliable applications of AI.

12. CRediT authorship contribution statement

Md Raisul Kibria: Conceptualization, Methodology, Investigation, Visualization, Writing - original draft. **Sébastien Lafond:** Supervision, Conceptualization, Methodology, Writing - review & editing. **Janan Arslan:** Supervision, Methodology, Writing - review & editing.

13. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM computing surveys (CSUR) 51 (5) (2018) 1–42.

[2] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information fusion 58 (2020) 82–115.

[3] N. Burkart, M. F. Huber, A survey on the explainability of supervised machine learning, Journal of Artificial Intelligence Research 70 (2021) 245–317.

[4] G. Yang, Q. Ye, J. Xia, Unbox the black-box for the medical explainable ai via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond, Information Fusion 77 (2022) 29–52.

[5] L. Nannini, A. Balayn, A. L. Smith, Explainability in ai policies: A critical review of communications, reports, regulations, and standards in the eu, us, and uk, in: Proceedings of the 2023 ACM conference on fairness, accountability, and transparency, 2023, pp. 1198–1212.

[6] J. Chun, C. S. de Witt, K. Elkins, Comparative global ai regulation: Policy perspectives from the eu, china, and the us, arXiv preprint arXiv:2410.21279 (2024).

[7] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer, 2014, pp. 740–755.

[8] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S. Bowman, GLUE: A multi-task benchmark and analysis platform for natural language understanding, in: T. Linzen, G. Chrupała, A. Alishahi (Eds.), Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 353–355. doi:10.18653/v1/W18-5446.
URL https://aclanthology.org/W18-5446

[9] O. Pelka, S. Koitka, J. Rückert, F. Nensa, C. M. Friedrich, Radiology objects in context (roco): a multimodal image dataset, in: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3, Springer, 2018, pp. 180–189.

[10] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, D. Parikh, Vqa: Visual question answering, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 2425–2433.

[11] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al., Visual genome: Connecting language and vision using crowdsourced dense image annotations, International journal of computer vision 123 (2017) 32–73.

[12] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R. G. Mark, Mimic-iii, a freely accessible critical care database, Scientific data 3 (1) (2016) 1–9.

[13] D. Saxena, J. Cao, Generative adversarial networks (gans) challenges, solutions, and future directions, ACM Computing Surveys (CSUR) 54 (3) (2021) 1–42.

[14] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, L. Farhan, Review of deep learning: concepts, cnn architectures, challenges, applications, future directions, Journal of big Data 8 (2021) 1–74.

[15] Y. Liu, P. Li, X. Hu, Combining context-relevant features with multi-stage attention network for short text classification, Computer Speech & Language 71 (2022) 101268.

[16] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473 (2014).

[17] A. Vaswani, Attention is all you need, Advances in Neural Information Processing Systems (2017).

[18] A. Dosovitskiy, An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[19] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, H. Hu, Video swin transformer, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 3202–3211.

[20] P. Xu, X. Zhu, D. A. Clifton, Multimodal learning with transformers: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (10) (2023) 12113–12132.

[21] S. S. Sengar, A. B. Hasan, S. Kumar, F. Carroll, Generative artificial intelligence: A systematic review and applications, arXiv preprint arXiv:2405.11029 (2024).

[22] S. Abnar, W. Zuidema, Quantifying attention flow in transformers, arXiv preprint arXiv:2005.00928 (2020).

[23] Y. Qiang, D. Pan, C. Li, X. Li, R. Jang, D. Zhu, AttCAT: Explaining Transformers via Attentive Class Activation Tokens.

[24] L. Parcalabescu, A. Frank, Mm-shap: A performance-agnostic metric for measuring multimodal contributions in vision and language models & tasks, arXiv preprint arXiv:2212.08158 (2022).

[25] N. Rodis, C. Sardianos, P. Radoglou-Grammatikis, P. Sarigiannidis, I. Varlamis, G. T. Papadopoulos, Multimodal explainable artificial intelligence: A comprehensive review of methodological advances and future research directions, IEEE Access (2024).

[26] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, arXiv preprint arXiv:1702.08608 (2017).

[27] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining explanations: An overview of interpretability of machine learning, in: 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA), IEEE, 2018, pp. 80–89.

[28] P. Fantozzi, M. Naldi, The explainability of transformers: Current status and directions, Computers 13 (4) (2024) 92.

[29] S. Mohseni, N. Zarei, E. D. Ragan, A multidisciplinary survey and framework for design and evaluation of explainable ai systems, ACM Transactions on Interactive Intelligent Systems (TiiS) 11 (3-4) (2021) 1–45.

[30] B. Kitchenham, Procedures for performing systematic reviews, Keele, UK, Keele University 33 (2004) (2004) 1–26.

[31] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, Preferred reporting items for systematic reviews and meta-analyses: the prisma statement, Bmj 339 (2009).

[32] S. Altmäe, A. Sola-Leyva, A. Salumets, Artificial intelligence in scientific writing: a friend or a foe?, Reproductive BioMedicine Online 47 (1) (2023) 3–9.

[33] J. Dagdelen, A. Dunn, S. Lee, N. Walker, A. S. Rosen, G. Ceder, K. A. Persson, A. Jain, Structured information extraction from scientific text with large language models, Nature Communications 15 (1) (2024) 1418.

[34] K. R. Felizardo, M. S. Lima, A. Deizepe, T. U. Conte, I. Steinmacher, Chatgpt application in systematic literature reviews in software engineering: an evaluation of its accuracy to support the selection activity, in: Proceedings of the 18th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2024, pp. 25–36.

[35] A. Huotala, M. Kuutila, P. Ralph, M. Mäntylä, The promise and challenges of using llms to accelerate the screening process of systematic reviews, in: Proceedings of the 28th International Conference on Evaluation and Assessment in Software Engineering, 2024, pp. 262–271.

[36] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).

[37] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Advances in neural information processing systems 35 (2022) 24824–24837.

[38] C. Wohlin, Guidelines for snowballing in systematic literature studies and a replication in software engineering, in: Proceedings of the 18th international conference on evaluation and assessment in software engineering, 2014, pp. 1–10.

[39] H. Chefer, S. Gur, L. Wolf, Generic Attention-model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Montreal, QC, Canada, 2021, pp. 387–396. doi:10.1109/ICCV48922.2021.00045.
URL https://ieeexplore.ieee.org/document/9710570/

[40] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. Van Keulen, C. Seifert, From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai, ACM Computing Surveys 55 (13s) (2023) 1–42.

[41] Y. Yang, L. Jiao, F. Liu, X. Liu, L. Li, P. Chen, S. Yang, An Explainable Spatial–Frequency Multiscale Transformer for Remote Sensing Scene Classification, IEEE Transactions on Geoscience and Remote Sensing 61 (2023) 1–15. doi:10.1109/TGRS.2023.3265361.
URL https://ieeexplore.ieee.org/document/10097579/

[42] Y. Huang, A. Jia, X. Zhang, J. Zhang, Generic Attention-model Explainability by Weighted Relevance Accumulation, in: ACM Multimedia Asia 2023, ACM, Tainan Taiwan, 2023, pp. 1–7. doi:10.1145/3595916.3626437.
URL https://dl.acm.org/doi/10.1145/3595916.3626437

[43] Y. Guo, F. Cai, H. Chen, C. Chen, X. Zhang, M. Zhang, An Explainable Recommendation Method based on Diffusion Model, in: 2023 9th International Conference on Big Data and Information Analytics (BigDIA), IEEE, Haikou, China, 2023, pp. 802–806. doi:10.1109/BigDIA60676.2023.10429319.
URL https://ieeexplore.ieee.org/document/10429319/

[44] Z. Liang, Y. Zhao, M. Surdeanu, Using the Hammer only on Nails: A Hybrid Method for Representation-Based Evidence Retrieval for Question Answering, in: D. Hiemstra, M.-F. Moens, J. Mothe, R. Perego, M. Potthast, F. Sebastiani (Eds.), Advances in Information Retrieval, Vol. 12656, Springer International Publishing, Cham, 2021, pp. 327–341, series Title: Lecture Notes in Computer Science. doi:10.1007/978-3-030-72113-8_22.
URL https://link.springer.com/10.1007/978-3-030-72113-8_22

[45] H. Wang, Y. Gao, Y. Bai, M. Lapata, H. Huang, Exploring Explainable Selection to Control Abstractive Summarization, Proceedings of the AAAI Conference on Artificial Intelligence 35 (15) (2021) 13933–13941. doi:10.1609/aaai.v35i15.17641.
URL https://ojs.aaai.org/index.php/AAAI/article/view/17641

[46] I. Malkiel, D. Ginzburg, O. Barkan, A. Caciularu, J. Weill, N. Koenigstein, Interpreting BERT-based Text Similarity via Activation and Saliency Maps, arXiv:2208.06612 [cs] (Aug. 2022). doi:10.48550/arXiv.2208.06612.
URL http://arxiv.org/abs/2208.06612

[47] J. Ferrando, M. R. Costa-jussà, Attention Weights in Transformer NMT Fail Aligning Words Between Sequences but Largely Explain Model Predictions, arXiv:2109.05853 [cs] (Sep. 2021). doi:10.48550/arXiv.2109.05853.
URL http://arxiv.org/abs/2109.05853

[48] K. Zhang, L. Li, Explainable multimodal trajectory prediction using attention models, Transportation Research Part C: Emerging Technologies 143 (2022) 103829. doi:10.1016/j.trc.2022.103829.
URL https://linkinghub.elsevier.com/retrieve/pii/S0968090X22002509

[49] M. Treviso, N. M. Guerreiro, R. Rei, A. F. T. Martins, IST-Unbabel 2021 Submission for the Explainable Quality Estimation Shared Task, in: Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 133–145. doi:10.18653/v1/2021.eval4nlp-1.14.
URL https://aclanthology.org/2021.eval4nlp-1.14

[50] S. Wang, Q. Zeng, W. Ni, C. Cheng, Y. Wang, ODP-Transformer: Interpretation of pest classification results using image caption generation techniques, Computers and Electronics in Agriculture 209 (2023) 107863. doi:10.1016/j.compag.2023.107863.
URL https://linkinghub.elsevier.com/retrieve/pii/S016816992300251X

[51] J. Sun, S. Wang, J. Zhang, C. Zong, Neural Encoding and Decoding With Distributed Sentence Representations, IEEE Transactions on Neural Networks and Learning Systems 32 (2) (2021) 589–603. doi:10.1109/TNNLS.2020.3027595.
URL https://ieeexplore.ieee.org/document/9223750/

[52] L. Yu, W. Xiang, J. Fang, Y.-P. P. Chen, L. Chi, eX-ViT: A Novel explainable vision transformer for weakly supervised semantic segmentation, Pattern Recognition 142 (2023) 109666. doi:10.1016/j.patcog.2023.109666.
URL https://linkinghub.elsevier.com/retrieve/pii/S0031320323003679

[53] M. Parelli, D. Mallis, M. Diomataris, V. Pitsikalis, Interpretable Visual Question Answering via Reasoning Supervision, arXiv:2309.03726 [cs] (Sep. 2023). doi:10.48550/arXiv.2309.03726.
URL http://arxiv.org/abs/2309.03726

[54] E. Aflalo, M. Du, S.-Y. Tseng, Y. Liu, C. Wu, N. Duan, V. Lal, VL-InterpreT: An Interactive Visualization Tool for Interpreting Vision-Language Transformers, arXiv:2203.17247 [cs] (Aug. 2022). doi:10.48550/arXiv.2203.17247.
URL http://arxiv.org/abs/2203.17247

[55] S. Katz, Y. Belinkov, VISIT: Visualizing and Interpreting the Semantic Information Flow of Transformers, arXiv:2305.13417 [cs] (Nov. 2023). doi:10.48550/arXiv.2305.13417.
URL http://arxiv.org/abs/2305.13417

[56] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, J. Gao, Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks, arXiv:2004.06165 [cs] (Jul. 2020). doi:10.48550/arXiv.2004.06165.
URL http://arxiv.org/abs/2004.06165

[57] R. K. Kandukuri, J. Achterhold, M. Moeller, J. Stueckler, Physical Representation Learning and Parameter Identification from Video Using Differentiable Physics, International Journal of Computer Vision 130 (1) (2022) 3–16. doi:10.1007/s11263-021-01493-5.
URL https://link.springer.com/10.1007/s11263-021-01493-5

[58] W. Sun, C. Wang, H. Wu, Y. Miao, H. Zhu, W. Guo, J. Li, DFYOLOv5m-M2transformer: Interpretation of vegetable disease recognition results using image dense captioning techniques, Computers and Electronics in Agriculture 215 (2023) 108460. doi:10.1016/j.compag.2023.108460.
URL https://linkinghub.elsevier.com/retrieve/pii/S0168169923008487

[59] M. Rigotti, C. Miksovic, I. Giurgiu, T. Gschwind, P. Scotton, ATTENTION-BASED INTERPRETABILITY WITH CONCEPT TRANSFORMERS (2022).

[60] Y. Heo, S. Kang, J. Seo, Natural-Language-Driven Multimodal Representation Learning for Audio-Visual Scene-Aware Dialog System, Sensors 23 (18) (2023) 7875. doi:10.3390/s23187875.
URL https://www.mdpi.com/1424-8220/23/18/7875

[61] J. Dong, S. Chen, M. Miralinaghi, T. Chen, P. Li, S. Labi, Why did the AI make that decision? Towards an explainable artificial intelligence (XAI) for autonomous driving systems, Transportation Research Part C: Emerging Technologies 156 (2023) 104358. doi:10.1016/j.trc.2023.104358.
URL https://linkinghub.elsevier.com/retrieve/pii/S0968090X23003480

[62] A. H. Mohammadkhani, C. Tantithamthavorn, H. Hemmatif, Explaining Transformer-based Code Models: What Do They Learn? When They Do Not Work?, in: 2023 IEEE 23rd International Working Conference on Source Code Analysis and Manipulation (SCAM), IEEE, Bogotá, Colombia, 2023, pp. 96–106.

doi:10.1109/SCAM59687.2023.00020.
URL https://ieeexplore.ieee.org/document/10356671/

[63] R. Buoy, M. Iwamura, S. Srun, K. Kise, Explainable Connectionist-Temporal-Classification-Based Scene Text Recognition, Journal of Imaging 9 (11) (2023) 248. doi:10.3390/jimaging9110248.
URL https://www.mdpi.com/2313-433X/9/11/248

[64] M. Z. Boito, A. Villavicencio, L. Besacier, Investigating alignment interpretability for low-resource NMT, Machine Translation 34 (4) (2020) 305–323. doi:10.1007/s10590-020-09254-w.
URL http://link.springer.com/10.1007/s10590-020-09254-w

[65] T. Chen, S. Liu, Z. Chen, W. Hu, D. Chen, Y. Wang, Q. Lyu, C. X. Le, W. Wang, Faster, Stronger, and More Interpretable: Massive Transformer Architectures for Vision-Language Tasks, Advances in Artificial Intelligence and Machine Learning 03 (03) (2023) 1369–1388. doi:10.54364/AAIML.2023.1181.
URL https://www.oajaiml.com/uploads/archivepdf/50081181.pdf

[66] N. Jaitly, Q. V. Le, O. Vinyals, I. Sutskever, D. Sussillo, S. Bengio, An online sequence-to-sequence model using partial conditioning, Advances in neural information processing systems 29 (2016).

[67] R. Prabhavalkar, T. N. Sainath, B. Li, K. Rao, N. Jaitly, An analysis of" attention" in sequence-to-sequence models., in: Interspeech, 2017, pp. 3702–3706.

[68] Y. Meng, W. Speier, M. K. Ong, C. W. Arnold, Bidirectional Representation Learning From Transformers Using Multimodal Electronic Health Record Data to Predict Depression, IEEE Journal of Biomedical and Health Informatics 25 (8) (2021) 3121–3129. doi:10.1109/JBHI.2021.3063721.
URL https://ieeexplore.ieee.org/document/9369833/

[69] D. Lee, C. H. Suh, J. Kim, W. Jung, C. Park, K.-H. Jung, S. T. Kong, W. H. Shim, H. Heo, S. J. Kim, Augmenting Magnetic Resonance Imaging with Tabular Features for Enhanced and Interpretable Medial Temporal Lobe Atrophy Prediction, in: A. Abdulkadir, D. R. Bathula, N. C. Dvornek, M. Habes, S. M. Kia, V. Kumar, T. Wolfers (Eds.), Machine Learning in Clinical Neuroimaging, Vol. 13596, Springer Nature Switzerland, Cham, 2022, pp. 125–134, series Title: Lecture Notes in Computer Science. doi:10.1007/978-3-031-17899-3_13.
URL https://link.springer.com/10.1007/978-3-031-17899-3_13

[70] C. C. Ukwuoma, Z. Qin, M. Belal Bin Heyat, F. Akhtar, O. Bamisile, A. Y. Muaad, D. Addo, M. A. Al-antari, A hybrid explainable ensemble transformer encoder for pneumonia identification from chest X-ray images, Journal of Advanced Research 48 (2023) 191–211. doi:10.1016/j.jare.2022.08.021.
URL https://linkinghub.elsevier.com/retrieve/pii/S2090123222002028

[71] T. Chiewhawan, P. Vateekul, Explainable Deep Learning for Thai Stock Market Prediction Using Textual Representation and Technical Indicators, in: Proceedings of the 8th International Conference on Computer and Communications Management, ACM, Singapore Singapore, 2020, pp. 19–23. doi:10.1145/3411174.3411191.

[72] B. Wu, L. Wang, Y.-R. Zeng, Interpretable tourism demand forecasting with temporal fusion transformers amid COVID-19, Applied Intelligence 53 (11) (2023) 14493–14514. doi:10.1007/s10489-022-04254-0.
URL https://link.springer.com/10.1007/s10489-022-04254-0

[73] D. Wang, W. Li, X. Dong, H. Li, L. Hu, TFRegNCI: Interpretable Noncovalent Interaction Correction Multimodal Based on Transformer Encoder Fusion, Journal of Chemical Information and Modeling 63 (3) (2023) 782–793. doi:10.1021/acs.jcim.2c01283.
URL https://pubs.acs.org/doi/10.1021/acs.jcim.2c01283

[74] U. Naseem, M. Khushi, J. Kim, Vision-Language Transformer for Interpretable Pathology Visual Question Answering, IEEE Journal of Biomedical and Health Informatics 27 (4) (2023) 1681–1690. doi:10.1109/JBHI.2022.3163751.
URL https://ieeexplore.ieee.org/document/9745795/

[75] S. Xu, W. Zhang, F. Zhang, Multi-Granular BERT: An Interpretable Model Applicable to Internet-of-Thing devices, in: 2020 IEEE International Conference on Energy Internet (ICEI), IEEE, Sydney, NSW, Australia, 2020, pp. 134–139. doi:10.1109/ICEI49372.2020.00032.
URL https://ieeexplore.ieee.org/document/9270262/

[76] B. Janssens, L. Schetgen, M. Bogaert, M. Meire, D. Van Den Poel, 360 Degrees rumor detection: When explanations got some explaining to do, European Journal of Operational Research 317 (2) (2024) 366–381. doi:10.1016/j.ejor.2023.06.024.
URL https://linkinghub.elsevier.com/retrieve/pii/S0377221723004769

[77] P. Ding, Y. Wang, X. Zhang, X. Gao, G. Liu, B. Yu, DeepSTF: predicting transcription factor binding sites by interpretable deep neural networks combining sequence and shape, Briefings in Bioinformatics 24 (4) (2023) bbad231. doi:10.1093/bib/bbad231.
URL https://academic.oup.com/bib/article/doi/10.1093/bib/bbad231/7199560

[78] M. Feucht, Z. Wu, S. Althammer, V. Tresp, Description-based Label Attention Classifier for Explainable ICD-9 Classification, arXiv:2109.12026 [cs] (Sep. 2021). doi:10.48550/arXiv.2109.12026.
URL http://arxiv.org/abs/2109.12026

[79] P. Kumar, V. Kaushik, B. Raman, Towards the Explainability of Multimodal Speech Emotion Recognition, in: Interspeech 2021, ISCA, 2021, pp. 1748–1752. doi:10.21437/Interspeech.2021-1718.
URL https://www.isca-archive.org/interspeech_2021/kumar21d_interspeech.html

[80] F. Ullah, A. Alsirhani, M. M. Alshahrani, A. Alomari, H. Naeem, S. A. Shah, Explainable Malware Detection System Using Transformers-Based Transfer Learning and Multi-Model Visual Representation, Sensors 22 (18) (2022) 6766. doi:10.3390/s22186766.

[81] F. M. Zanzotto, A. Santilli, L. Ranaldi, D. Onorati, P. Tommasino, F. Fallucchi, KERMIT: Complementing Transformer Architectures with Encoders of Explicit Syntactic Interpretations, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 256–267. doi:10.18653/v1/2020.emnlp-main.18.
URL https://www.aclweb.org/anthology/2020.emnlp-main.18

[82] F. Xu, J. Liu, Q. Lin, Y. Pan, L. Zhang, Logiformer: A Two-Branch Graph Transformer Network for Interpretable Logical Reasoning, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, Madrid Spain, 2022, pp. 1055–1065. doi:10.1145/3477495.3532016.
URL https://dl.acm.org/doi/10.1145/3477495.3532016

[83] A. Jha, V. Rakesh, J. Chandrashekar, A. Samavedhi, C. K. Reddy, Supervised Contrastive Learning for Interpretable Long-Form Document Matching, ACM Transactions on Knowledge Discovery from Data 17 (2) (2023) 1–17. doi:10.1145/3542822.
URL https://dl.acm.org/doi/10.1145/3542822

[84] S. Che, J. Lu, C. Bao, C. Zhang, Y. Liu, Multiscale Time-Frequency Sparse Transformer Based on Partly Interpretable Method for Bearing Fault Diagnosis, Shock and Vibration 2023 (2023) 1–21. doi:10.1155/2023/1639287.
URL https://www.hindawi.com/journals/sv/2023/1639287/

[85] N. Ilinykh, S. Dobnik, What Does a Language-And-Vision Transformer See: The Impact of Semantic Information on Visual Representations, Frontiers in Artificial Intelligence 4 (2021) 767971. doi:10.3389/frai.2021.767971.
URL https://www.frontiersin.org/articles/10.3389/frai.2021.767971/full

[86] F. Lin, M. Li, D. Li, T. Hospedales, Y.-Z. Song, Y. Qi, Zero-Shot Everything Sketch-Based Image Retrieval, and in Explainable Style, arXiv:2303.14348 [cs] (Mar. 2023). doi:10.48550/arXiv.2303.14348.
URL http://arxiv.org/abs/2303.14348

[87] J. Ferrando, G. I. Gállego, B. Alastruey, C. Escolano, M. R. Costa-jussà, Towards Opening the Black Box of Neural Machine Translation: Source and Target Interpretations of the Transformer, arXiv:2205.11631 [cs] (Nov. 2022). doi:10.48550/arXiv.2205.11631.
URL http://arxiv.org/abs/2205.11631

[88] Y. Huang, M. Wang, Z. Zheng, M. Ma, X. Fei, L. Wei, H. Chen, Representation of time-varying and time-invariant EMR data and its application in modeling outcome prediction for heart failure patients, Journal of Biomedical Informatics 143 (2023) 104427. doi:10.1016/j.jbi.2023.104427.
URL https://linkinghub.elsevier.com/retrieve/pii/S153204642300148X

[89] Y. Zheng, P. Tang, W. Qiu, H. Wang, J. Guo, Z. Huang, A Novel Deep Learning Framework for Interpretable Drug-Target Interaction Prediction with Attention

and Multi-task Mechanism, in: X. Wang, M. L. Sapino, W.-S. Han, A. El Abbadi, G. Dobbie, Z. Feng, Y. Shao, H. Yin (Eds.), Database Systems for Advanced Applications, Vol. 13946, Springer Nature Switzerland, Cham, 2023, pp. 336–352, series Title: Lecture Notes in Computer Science. doi:10.1007/978-3-031-30678-5_26.

[90] K. Koyama, K. Hashimoto, C. Nagao, K. Mizuguchi, Attention network for predicting T-cell receptor–peptide binding can associate attention with interpretable protein structural properties, Frontiers in Bioinformatics 3 (2023) 1274599. doi:10.3389/fbinf.2023.1274599.
URL https://www.frontiersin.org/articles/10.3389/fbinf.2023.1274599/full

[91] P. Bhargava, Adaptive Transformers for Learning Multimodal Representations, arXiv:2005.07486 [cs] (Jul. 2020). doi:10.48550/arXiv.2005.07486.
URL http://arxiv.org/abs/2005.07486

[92] Y. Du, Z. Guan, W. Huang, X. Zhang, Q. Huang, A Case-based Channel Selection Method for EEG Emotion Recognition Using Interpretable Transformer Networks, in: Proceedings of the 2023 International Conference on Computer, Vision and Intelligent Technology, ACM, Chenzhou China, 2023, pp. 1–5. doi:10.1145/3627341.3630372.

[93] P. Kumar, B. Raman, A BERT based dual-channel explainable text emotion recognition system, Neural Networks 150 (2022) 392–407. doi:10.1016/j.neunet.2022.03.017.
URL https://linkinghub.elsevier.com/retrieve/pii/S0893608022000958

[94] X. Xiao, Y. Kong, R. Li, Z. Wang, H. Lu, Transformer with convolution and graph-node co-embedding: An accurate and interpretable vision backbone for predicting gene expressions from local histopathological image, Medical Image Analysis 91 (2024) 103040. doi:10.1016/j.media.2023.103040.
URL https://linkinghub.elsevier.com/retrieve/pii/S1361841523003006

[95] H. Tan, M. Bansal, Lxmert: Learning cross-modality encoder representations from transformers, arXiv preprint arXiv:1908.07490 (2019).

[96] G. M. Correia, V. Niculae, A. F. Martins, Adaptively sparse transformers, arXiv preprint arXiv:1909.00015 (2019).

[97] G. Sarti, N. Feldhus, L. Sickert, O. v. d. Wal, M. Nissim, A. Bisazza, Inseq: An Interpretability Toolkit for Sequence Generation Models, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), 2023, pp. 421–435, arXiv:2302.13942 [cs]. doi:10.18653/v1/2023.acl-demo.40.
URL http://arxiv.org/abs/2302.13942

[98] G. Vilone, L. Longo, Classification of explainable artificial intelligence methods through their output formats, Machine Learning and Knowledge Extraction 3 (3) (2021) 615–661.

[99] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, J. Zhu, Explainable ai: A brief survey on history, research areas, approaches and challenges, in: Natural language processing and Chinese computing: 8th cCF international conference, NLPCC 2019, dunhuang, China, October 9–14, 2019, proceedings, part II 8, Springer, 2019, pp. 563–574.

[100] S. Jain, B. C. Wallace, Attention is not explanation, arXiv preprint arXiv:1902.10186 (2019).

[101] J. Vig, A multiscale visualization of attention in the transformer model, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Florence, Italy, 2019, pp. 37–42. doi:10.18653/v1/P19-3007.
URL https://www.aclweb.org/anthology/P19-3007

[102] J. Ferrando, G. I. Gállego, M. R. Costa-Jussà, Measuring the mixing of contextual information in the transformer, arXiv preprint arXiv:2203.04212 (2022).

[103] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

[104] H. Chefer, S. Gur, L. Wolf, Transformer interpretability beyond attention visualization, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 782–791.

[105] A. Fan, E. Grave, A. Joulin, Reducing transformer depth on demand with structured dropout, arXiv preprint arXiv:1909.11556 (2019).

[106] M. Hartmann, D. Sonntag, A survey on improving nlp models with human explanations, arXiv preprint arXiv:2204.08892 (2022).

[107] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (8) (2019) 9.

[108] A. Hedström, L. Weber, D. Krakowczyk, D. Bareeva, F. Motzkus, W. Samek, S. Lapuschkin, M. M.-C. Höhne, Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond, Journal of Machine Learning Research 24 (34) (2023) 1–11.

[109] G. Vilone, L. Longo, Notions of explainability and evaluation approaches for explainable artificial intelligence, Information Fusion 76 (2021) 89–106.

[110] E. M. Voorhees, et al., The trec-8 question answering track report., in: Trec, Vol. 99, 1999, pp. 77–82.

[111] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[112] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.

[113] R. Vedantam, C. Lawrence Zitnick, D. Parikh, Cider: Consensus-based image description evaluation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4566–4575.

[114] P. Anderson, B. Fernando, M. Johnson, S. Gould, Spice: Semantic propositional image caption evaluation, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14, Springer, 2016, pp. 382–398.

[115] M. Denkowski, A. Lavie, Meteor universal: Language specific translation evaluation for any target language, in: Proceedings of the ninth workshop on statistical machine translation, 2014, pp. 376–380.

[116] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, C. L. Zitnick, Microsoft coco captions: Data collection and evaluation server, arXiv preprint arXiv:1504.00325 (2015).

[117] D. V. Carvalho, E. M. Pereira, J. S. Cardoso, Machine learning interpretability: A survey on methods and metrics, Electronics 8 (8) (2019) 832.

[118] K. Yoshino, Y.-N. Chen, P. Crook, S. Kottur, J. Li, B. Hedayatnia, S. Moon, Z. Fei, Z. Li, J. Zhang, et al., Overview of the tenth dialog system technology challenge: Dstc10, IEEE/ACM Transactions on Audio, Speech, and Language Processing 32 (2023) 765–778.

[119] Y. Liu, C. Wu, S.-y. Tseng, V. Lal, X. He, N. Duan, Kd-vlp: Improving end-to-end vision-and-language pretraining with object knowledge distillation, arXiv preprint arXiv:2109.10504 (2021).

[120] S. Serrano, N. A. Smith, Is attention interpretable?, arXiv preprint arXiv:1906.03731 (2019).

[121] B. Bai, J. Liang, G. Zhang, H. Li, K. Bai, F. Wang, Why attentions may not be interpretable?, in: Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining, 2021, pp. 25–34.

[122] L. Wenderoth, K. Hemker, N. Simidjievski, M. Jamnik, Measuring cross-modal interactions in multimodal models, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 39, 2025, pp. 21501–21509.

[123] J. P. Sheppard, D. Raposo, A. K. Churchland, Dynamic weighting of multisensory stimuli shapes decision-making in rats and humans, Journal of vision 13 (6) (2013) 4–4.

[124] S. R. Islam, W. Eberle, S. K. Ghafoor, Towards quantification of explainability in explainable artificial intelligence methods, in: The thirty-third international flairs conference, 2020.

[125] X. Liu, S. C. Rivera, D. Moher, M. J. Calvert, A. K. Denniston, H. Ashrafian, A. L. Beam, A.-W. Chan, G. S. Collins, A. D. J. Deeks, et al., Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the consort-ai extension, The Lancet Digital Health 2 (10) (2020) e537–e548.

[126] J. E. T. Taylor, G. W. Taylor, Artificial cognition: How experimental psychology can help generate explainable artificial intelligence, Psychonomic Bulletin & Review 28 (2) (2021) 454–475.

[127] W. F. Lawless, R. Mittu, D. Sofge, L. Hiatt, Artificial intelligence, autonomy, and human-machine teams—interdependence, context, and explainable ai, Ai Magazine 40 (3) (2019) 5–13.

[128] D. Wang, Q. Yang, A. Abdul, B. Y. Lim, Designing theory-driven user-centric explainable ai, in: Proceedings of the 2019 CHI conference on human factors in computing systems, 2019, pp. 1–15.

[129] T. Rohe, U. Noppeney, Cortical hierarchies perform bayesian causal inference in multisensory perception, PLoS biology 13 (2) (2015) e1002073.

[130] P. P. Liang, Y. Lyu, X. Fan, J. Tsaw, Y. Liu, S. Mo, D. Yogatama, L.-P. Morency, R. Salakhutdinov, High-modality multimodal transformer: Quantifying modality & interaction heterogeneity for high-modality representation learning, arXiv preprint arXiv:2203.01311 (2022).

[131] J. Jeong, K. Tian, A. Li, S. Hartung, S. Adithan, F. Behzadi, J. Calle, D. Osayande, M. Pohlen, P. Rajpurkar, Multimodal image-text matching improves retrieval-based chest x-ray report generation, in: Medical Imaging with Deep Learning, PMLR, 2024, pp. 978–990.

[132] C. Agarwal, Rethinking explainability in the era of multimodal ai, arXiv preprint arXiv:2506.13060 (2025).