CLEAR

Comprehensive Learning for Enhanced AI Responsiveness

D2.1 - SotA and Guidelines for Multimodal Dataset Creation, Collection, and Fusion

Submission date of deliverable: July 28, 2026

Edited by: Mohammad Loni (FormalTrust.ai), Silke Cleuren (Materialise NV), Hakan Kilinc (Orion), Selen Pehlivan Tort (VTT), Paul Kemppi (VTT), Sebnem Koken (Defne), Abbas Khan (RISE), Abdul Rauf (Test Scouts)

Project start date	Dec 1, 2025
Project duration	36 months
Project coordinator	Mehrdad Saadatmand, RISE Research Institutes of Sweden
Project number & call	24026 - ITEA Call 2024
Project website	https://itea4.org/project/clear.html

Contributing partners	WP2 partners
Version number	V1.0
Work package	WP2: Data collection; Aggregation and Fusion
Work package leader	Mohammad Loni, FormalTrust.ai
Dissemination level	Public

Description	This deliverable report collects the state-of-the-art at the start of the project for multimodal dataset creation, collection and fusion. This allows us to clearly know what the status was at the start of the project and how, throughout the project, we move beyond that. In addition, it will serve as a guide to set-up the project's datasets.

Executive Summary

The main objective of WP2 is to establish a solid and systematic data foundation for the CLEAR project by enabling the creation of high-quality, reliable, and diverse multi-modal datasets that support the development, fine-tuning, and evaluation of Large Language Models (LLMs) and Large Multimodal Models (LMMs) in industrial environments. WP2 addresses the challenges of integrating heterogeneous data sources, managing data scarcity and confidentiality, and enabling effective multimodal data fusion across a wide range of industrial domains. Specifically, WP2 aims to:

- Define state-of-the-art practices and guidelines for the creation, collection, and management of multi-modal datasets (This Deliverable),
- Enable structured aggregation and preparation of domain-specific and cross-domain data required by industrial use cases,
- Investigate and apply early, intermediate, late, and hybrid fusion strategies to combine heterogeneous data modalities,
- Address data confidentiality, privacy, and access constraints through anonymisation, controlled data sharing, and synthetic data generation,
- Provide robust, reusable datasets that serve as reliable inputs for model construction, fine-tuning, and pipeline development in WP3 and WP4.

This deliverable is the primary output of Task 2.1 and establishes the conceptual and methodological foundation for all data-related activities in CLEAR. It provides a comprehensive review of the state of the art and state of practice in multimodal data handling, with a particular focus on industrial AI applications involving unconventional and heterogeneous data types such as sensor data, time series, images, text, audio, video, and 3D data. The purpose of Task 2.1 and this deliverable can be summarised as follows:

- Review and analyse SotA and best practices for multi-modal dataset creation, annotation, and management in industrial AI systems,
- Identify challenges related to data heterogeneity, scale, alignment, quality, scarcity, and confidentiality,
- Define practical guidelines for collecting, organising, and documenting multi-modal datasets in a consistent and reusable manner,
- Provide methodological guidance on selecting and applying appropriate fusion strategies depending on data characteristics and use-case requirements,
- Establish common principles for ensuring data quality, traceability, and suitability for fine-tuning and evaluation of LLMs and LMMs.

CLEAR's industrial partners operate across diverse sectors, including railway and transportation, high-tech equipment, telecommunications, agriculture, industrial automation, additive manufacturing, and software-intensive systems. As a result, the guidelines presented in this document are designed to be domain-agnostic while remaining sufficiently concrete to support sector-specific dataset creation. The outcomes of D2.1 form a reference framework for subsequent WP2 tasks (T2.2 - T2.4) and directly support the downstream technical work in WP3 (model construction and fine-tuning) and WP4 (multi-modal pipeline development).

By establishing a shared understanding of best practices and methodological choices for multi-modal data handling, this deliverable ensures coherence, quality, and scalability of data-driven developments across the CLEAR project, thereby contributing to the reliability, trustworthiness, and industrial applicability of the resulting AI solutions.

Table of Contents

1. Industrial Multi-Modal Data Landscapes and Challenges

Industrial AI systems increasingly rely on the integration of data originating from a wide range of heterogeneous sources. Unlike consumer-oriented AI applications, industrial environments generate and manage data that differ significantly in structure, scale, dynamics, and semantic meaning. Understanding these characteristics and the associated challenges is essential for designing effective multi-modal datasets that can support reliable Large Language Models (LLMs) and Large Multimodal Models (LMMs), as targeted by the CLEAR project.

1.1. Characteristics of Industrial Data

Industrial data ecosystems are inherently multi-modal, combining information from diverse sources and sensing mechanisms. Typical data modalities include:

- textual data such as technical documentation, manuals, maintenance logs, and incident reports;
- structured data from databases and enterprise systems;
- sensor and time-series data generated by machines, production lines, and monitoring systems;
- visual data such as images and videos captured by cameras or inspection devices;
- audio data from operational environments; three-dimensional data including CAD files, meshes, and point clouds;
- and geospatial data derived from satellite imagery or positioning systems.

Each modality exhibits distinct data representations, noise characteristics, temporal resolution, and semantic structure, making unified handling and integration non-trivial.

This diversity directly motivates CLEAR Technological Innovation 1 (Flexible Multimodal Data Fusion), which targets adaptive fusion frameworks capable of preserving modality-specific characteristics while enabling effective cross-modal alignment. It also underpins Innovation 5 (LMM-Enhanced Multimodal Interpretation using Digital Twins), where heterogeneous data streams are interpreted in a shared contextual framework.

A further distinguishing aspect of industrial data is the coexistence of structured and unstructured data. While structured data, such as sensor readings or configuration parameters, can be readily processed using conventional data pipelines, unstructured data such as free-text reports, images, or videos require advanced preprocessing, representation learning, and semantic alignment. Addressing this heterogeneity is essential for enabling Innovation 6 (Composable Large Multimodal AI Architecture), where datasets must support modular and interchangeable processing components.

Industrial data can also be categorised as streaming or static. Streaming data, such as real-time sensor measurements or event logs, arrive continuously and often under strict latency constraints, whereas static data, such as manuals, design files, or historical maintenance records, are relatively stable but may be updated sporadically. Combining streaming and static data introduces challenges related to temporal alignment, versioning, and synchronisation, which are directly addressed by Innovation 4 (AIOps/MLOps for Continuous AI Model Management) and influence dataset design choices in this work package.

1.2. Cross-Domain Data Integration Challenges

Across industrial domains, data is typically distributed over multiple silos and legacy systems, reflecting long-standing organisational and technological evolution. These silos differ in data

formats, metadata conventions, storage technologies, and access policies. Extracting and harmonising data across such systems is a prerequisite for multimodal learning and directly supports Innovation 6, which relies on composable pipelines capable of integrating heterogeneous data sources.

Another major challenge arises from asynchronous and misaligned data sources. Different modalities are generated at different temporal and spatial resolutions and often lack explicit synchronisation. For example, high-frequency sensor streams must be aligned with irregularly generated maintenance logs or operator annotations. Addressing these challenges is essential for Innovation 1, which introduces intermediate and adaptive fusion mechanisms, and for Innovation 5, where Digital Twins rely on coherent temporal and semantic alignment across modalities.

Industrial datasets also commonly exhibit long-tail distributions, where a limited number of frequent events dominate the data, while rare but critical events, such as faults, anomalies, or safety-related incidents, are sparsely represented. These rare events are of high importance for diagnostics and decision-making but are difficult to capture in sufficient quantity. This challenge directly motivates Innovation 3 (Self-Supervised Learning Models in Data Scarcity Environments) and the synthetic data generation activities described in WP2.

1.3. Industrial Constraints and Requirements

Beyond technical challenges, industrial data handling is constrained by confidentiality, intellectual property protection, and data ownership requirements. Many datasets contain sensitive operational or commercial information that cannot be freely shared across organisational boundaries. Dataset creation workflows must therefore support anonymisation, controlled access, and the use of synthetic or representative data, providing a foundation for Innovation 2 (Explainability and Reasoning in Large Multimodal Pipelines) and Innovation 4, where secure and controlled data access is essential.

A further constraint is the limited availability of labelled data. Labelling industrial data often requires specialised domain expertise and manual effort, making it costly and time-consuming. This limitation reinforces the importance of Innovation 3, which focuses on self-supervised learning and synthetic data augmentation, and directly influences the dataset design guidelines presented in this deliverable.

Finally, many industrial applications operate in safety-critical and regulated environments, where incorrect or non-transparent AI outputs can lead to safety risks, financial losses, or regulatory non-compliance. As a result, datasets must support traceability, validation, and explainability, enabling downstream AI systems to justify their outputs. This requirement is tightly coupled with Innovation 2 and Innovation 5, where reasoning, provenance, and contextual understanding are core design principles.

Together, these characteristics and constraints highlight the complexity of industrial multi-modal data landscapes and justify the need for systematic guidelines for dataset creation, collection, and fusion, as developed in the remainder of this deliverable.

2. State of the Art in Multimodal Dataset Creation

2.1. Sensor-based Data Collection

Sensor-based data collection is a central theme in (Conti, Martínez Madrid and Seepold), where multiple chapters describe how multimodal biometric datasets are acquired using wearable devices, embedded sensors, and IoT systems. The book covers physiological sensing, cardiac monitoring, gait analysis, and smart environment systems, illustrating how real-world biometric data is collected across different application domains (Conti, Martínez Madrid and Seepold).For example, the SAFESA system collects physiological signals such as blood pressure, ECG, glucose levels, body temperature, and oxygen saturation using wearable sensors connected to smartphones, while also integrating environmental data like temperature and humidity for health analysis. Other contributions include ECG-based stress detection using controlled experimental protocols, low-cost photoplethysmogram (PPG) heart rate monitoring, and combined EMG/ECG sensor nodes for multimodal signal acquisition. Additionally, gait and motion data are collected using surface EMG systems, wearable accelerometers, and sports monitoring devices, while IoT-based systems extend data collection to smart environments such as food tracking shelves and clothing management systems (Conti et al., 2016). Across these systems, a common challenge is the presence of noise and motion artifacts in sensor signals, which affects data quality and reliability. The studies collectively highlight the complexity of real-world biometric data collection and the need for robust sensing, filtering, and experimental design strategies (Conti, Martínez Madrid and Seepold).

(Liu, Wu and Kumar) propose a sensor-based IoT data collection framework designed for intelligent marine economy applications. The study explores how sensor networks and IoT technologies can be integrated with underwater systems and robotic platforms to collect real-time environmental and operational data in marine ranching environments. The system uses distributed sensors and IoT communication technologies to gather and transmit data related to marine conditions and industrial activity. This data is then used for economic analysis and decision-making, aiming to improve the efficiency and sustainability of marine resource management. The authors highlight that combining sensor systems with IoT infrastructure enables continuous monitoring, large-scale data acquisition, and better support for industrial innovation in marine environments. However, the study also emphasises the importance of system integration and scalability for real-world deployment.

(Alam, Surani and Das) provide a review of sensor-based data collection in digital phenotyping, where wearable devices and smartphones are used to continuously collect behavioural and physiological data from individuals. This approach is increasingly applied in mental health research to detect early signs of symptom changes and support personalised interventions. The paper highlights that sensor-based systems can capture real-time signals such as heart rate, activity levels, and behavioural patterns, enabling large-scale monitoring of human behaviour outside clinical settings. However, the authors identify several challenges, including inconsistent data quality, lack of standardisation across devices, and interoperability issues between platforms. To address these limitations, the study proposes standardisation strategies, such as universal data collection protocols, open APIs, improved cross-platform compatibility, and collaboration between researchers and industry. The authors also emphasise the importance of user-centred and culturally sensitive design to improve data reliability and inclusiveness in digital health applications.

(Zhu and Brilakis) review different optical sensor-based techniques used to collect spatial data for civil infrastructure modelling, such as in construction, maintenance, and rehabilitation of buildings and infrastructure. The paper explains that multiple sensing methods exist for capturing spatial information, but none of them is universally optimal. The authors compare these techniques based on key factors including accuracy, automation level, cost, portability, and data retrieval efficiency. Their goal is to help practitioners select the most suitable sensor technology depending on specific project requirements rather than relying on a one-size-fits-all solution. Overall, the study

emphasises that infrastructure modelling relies heavily on sensor-driven data collection systems, and the choice of sensing method directly affects the quality and usability of the resulting spatial datasets.

2.2. Human-generated data

(Ahsen, Ayvaci and Raghunathan) study how machine learning systems behave when trained on human-generated data, particularly in the context of breast cancer diagnosis. The authors focus on a decision-support setting where a physician combines patient clinical-risk information with radiologist mammogram assessments to make follow-up recommendations. The paper highlights that radiologist evaluations can be influenced by bias from clinical information, meaning the input data itself is not fully objective. This bias is then inherited by machine learning models trained on such data, which can reduce prediction accuracy and lead to suboptimal clinical decisions. To address this, the authors propose a bias-aware classification algorithm that adjusts the weight of different input features depending on the level of bias and error in the data. The results show that accounting for bias can significantly improve decision quality and patient outcomes, including expected life years and diagnostic accuracy. Overall, the study demonstrates that human-generated data introduces systematic bias into machine learning systems, and that explicitly modelling this bias can improve algorithmic performance in healthcare applications.

(J. Park) explores machine learning methods for analysing human-generated sequential and temporal data in two domains: education and healthcare. In the education domain, the study uses student clickstream data to detect behavioural changes during online learning and applies probabilistic clustering to identify patterns such as procrastination and time-management behaviours, which are linked to course outcomes. In the medical domain, the dissertation analyses doctor–patient conversation transcripts to extract structured information using machine learning models such as logistic regression, support vector machines, GRUs, and sequence models like HMMs and CRFs. The study also examines how incorporating sequential context improves topic classification accuracy and evaluates the impact of automatic speech recognition errors on model performance. Additionally, it investigates emotional valence prediction at the utterance level in medical dialogs. Overall, the research demonstrates that incorporating temporal structure in human-generated data significantly improves predictive performance, while also highlighting challenges such as noise introduced by transcription systems.

(Villalobos, Ho and Sevilla) examine the limits of scaling large language models (LLMs) based on the availability of human-generated text data. The study estimates that the demand for training data is increasing rapidly and may reach the total supply of publicly available human-written text between 2026 and 2032. This suggests that current trends in LLM development could soon face a “data bottleneck,” where further performance improvements cannot rely solely on human-generated datasets. To address this limitation, the authors discuss alternative approaches such as synthetic data generation, transfer learning from data-rich domains, and improving data efficiency. Overall, the paper highlights a major constraint in AI development and emphasises the need for new strategies beyond traditional human-created datasets.

(Chen, Tang and Fan) explore how human-generated data pipelines and machine-generated (AutoML) pipelines can be combined to improve data preparation for machine learning tasks. Human-generated pipelines are typically based on expert experience and heuristic decisions, while machine-generated pipelines are more systematic and search-optimised. The authors propose a framework called HAIPipe, which integrates both approaches to create a hybrid pipeline that improves performance over using either method alone. The study shows that combining human expertise with automated optimisation leads to better results in real-world machine learning workflows. This highlights the importance of leveraging both human knowledge and machine efficiency in data preparation.

2.3. Crowdsourced and Field-Collected Data

(Dabbaghchian, Matarazzo and Sadeghi Eshkevari) studied the use of crowdsourced smartphone-vehicle trip (SVT) data for bridge structural health monitoring. The approach uses smartphone accelerometer and GPS data collected from vehicles crossing a bridge to estimate bridge modal frequencies and mode shapes, offering a cost-effective alternative to traditional sensor-based systems (Dabbaghchian, Matarazzo and Sadeghi Eshkevari).The study highlights that SVT data is affected by noise from sensors, vehicle motion, and road conditions, which reduces data quality. To address this, the authors propose a quality metric and a CNN-based model to evaluate and filter low-quality trips. Using over 900 field-collected trips from the Cadore Viaduct in Italy, the method improved modal estimation accuracy, increasing the Modal Assurance Criterion (MAC) from 0.80 to 0.97 after data filtering (Dabbaghchian, Matarazzo and Sadeghi Eshkevari). (Schepaschenko, See and Lesiv) develop a global hybrid forest mask by integrating multiple data sources, including remote sensing products, FAO forest statistics, and crowdsourced validation data collected through the Geo-Wiki platform. The main goal is to resolve inconsistencies between existing forest maps, which often disagree spatially and lack alignment with official forest statistics. The authors apply a geographically weighted regression (GWR) approach to combine eight different forest datasets into a unified global forest cover map at 1 km resolution. Crowdsourced data are used both for model training and independent validation, improving the reliability of the final product. Results show that the hybrid approach significantly improves accuracy, with the best-performing map achieving around 93% accuracy, outperforming individual input datasets. The study demonstrates that combining crowdsourced observations with remote sensing and statistical data can produce more accurate and consistent global environmental datasets.

2.4. Dataset Scope and Representativeness

(Carlesso, Pizzol and Marcomini) proposed a Data Quality Assessment (DQA) framework for aggregated Life Cycle Inventory (LCI) datasets used in life cycle assessment. The study addresses inconsistencies in environmental impact results caused by variations in dataset quality and representativeness. The authors evaluate 101 LCI datasets from the Ecoinvent and GaBi databases using three key indicators: technological, geographical, and time-related representativeness. Results show that dataset quality varies across these dimensions, highlighting the importance of structured metadata and standardised evaluation methods to improve reliability and transparency in environmental assessments. The framework can be extended to other sectors and dataset types to support more consistent data quality evaluation (Carlesso, Pizzol and Marcomini).

(Di Girolamo, Walters and Gildea) examined the completeness and representativeness of the English Cancer Waiting Times (CWT) dataset by linking it with national cancer registry and hospital records for patients diagnosed with colorectal, lung, and ovarian cancer (2009–2013). The study found that the dataset did not include all eligible patients, with coverage ranging from 76% to 82% depending on cancer type. Missing records were associated with specific patient characteristics, indicating systematic selection bias. The authors conclude that incomplete dataset coverage can affect research validity and should be considered when using administrative health datasets.

(Zhang, Cai and Qiu) provide a comprehensive survey of deep learning-based Action Quality Assessment (AQA) methods, which aim to evaluate the quality of human actions from video data. Unlike traditional action recognition, AQA focuses on assigning performance scores, making it useful in domains such as sports analysis, rehabilitation, and skill evaluation. The survey categorises existing approaches based on data modalities, learning strategies, and evaluation levels, including single-modal and multimodal methods that combine visual, audio, and textual information. It also reviews commonly used benchmark datasets and analyses their scope, representativeness, and annotation quality, highlighting how dataset design impacts model performance and generalisation. The authors identify key challenges such as limited dataset

generalisation, ambiguity in scoring labels, and interpretability issues, and suggest future research directions to improve dataset quality and model robustness in AQA systems.
(Van der Velde, Benninga and Retsios) present a 12-year dataset of soil moisture and soil temperature measurements collected from a monitoring network in Twente, Netherlands. The dataset includes observations from 20 stations measuring soil conditions at multiple depths (5, 10, 20, 40, and 80 cm), along with field campaign measurements and external supporting datasets such as land use, soil type, elevation, and meteorological data. The study describes the design of the monitoring network, sampling strategy, calibration methods, and quality control procedures used to ensure data reliability. A key focus is evaluating the spatial representativeness of the monitoring stations by comparing station measurements with field-averaged soil moisture values. Results show generally reasonable agreement, although errors are observed, particularly in grassland areas and after heavy rainfall events. Overall, the paper emphasises the importance of careful sampling design and representativeness assessment in long-term environmental monitoring datasets to ensure reliable use in climate, hydrology, and land-surface studies.
(Houck) introduces SIERA-ex, a framework designed to assess and improve the representativeness of ex situ conservation datasets for single-island endemic plant species in Kaua‘i. The study addresses the challenge that conservation repositories often hold incomplete or unevenly distributed genetic samples, which can limit biodiversity preservation efforts. The framework uses geographic and ecological coverage as proxies for genetic diversity, especially when direct genetic distance data is unavailable. It incorporates spatial resolution across multiple ecological and cultural layers, including watersheds and Hawaiian place-based regions, to evaluate how well conservation collections represent natural populations. A case study involving nine endemic Kaua‘i plant species demonstrates how the method identifies underrepresented subpopulations and prioritises them for conservation action. The results show that limiting scope and focusing on high-priority, high-threat taxa improves the effectiveness of conservation planning and resource allocation. Overall, the study highlights the importance of sampling strategy and representativeness in biodiversity datasets, particularly for conservation planning and ex situ collections.

2.5. Handling modality imbalance

(Alrayzah) investigates multimodal sentiment analysis, where systems learn human emotions by combining different data types such as text, audio, and visual signals. A key challenge in this area is modality imbalance, where one type of data (e.g., text) may dominate or be more reliable than others, leading to poor model performance and unstable learning. To address this, the study proposes a tri-modal cross-attention framework with adaptive gating mechanisms. This approach improves how the model integrates and balances information across modalities, allowing it to better capture relationships between text, audio, and visual inputs. Experiments on benchmark datasets (CMU-MOSI, CMU-MOSEI, and SIMS) show improved accuracy and F1-scores compared to existing methods. The results confirm that combining cross-modal attention with adaptive gating helps create more robust and balanced multimodal learning systems.
(Fujinami, Takuno and Sato) propose a machine learning–based behaviour recognition system for laying hens using wearable inertial sensors. The study focuses on monitoring animal behaviour to improve welfare in farming systems. The system collects continuous sensor data (such as acceleration and angular velocity) and uses supervised machine learning to classify 12 different hen behaviours. A key focus of the paper is the design of the full processing pipeline, including classifier choice, sampling frequency, window size, and data imbalance handling, since some behaviours appear more frequently than others. The authors show that these design choices significantly affect model performance and the reliability of behaviour recognition. Overall, the study demonstrates how sensor-based data and machine learning can be used to monitor animal welfare, while also highlighting the importance of handling imbalanced datasets in real-world applications.

(Tian, Yang and Zhu) investigate the problem of extreme modality imbalance in RGB-Thermal object detection, where one data modality (RGB or thermal infrared) becomes degraded or missing due to real-world conditions such as environmental interference or sensor failure. This imbalance can disrupt model training, reduce convergence stability, and lead to out-of-distribution (OOD) behaviour during testing. To address these issues, the authors propose a base-and-auxiliary detector framework in which a base detector is trained on high-quality paired data, while an auxiliary detector learns from degraded or imbalanced samples. A modality interaction module is introduced to dynamically adjust the contribution of each modality based on its quality, ensuring more reliable fusion. In addition, modality pseudo-degradation is used during training to simulate realistic imbalance scenarios and improve robustness. Experimental results show that the proposed approach improves detection performance under severe modality imbalance conditions and reduces missing-rate issues, demonstrating stronger generalisation in challenging real-world environments (Tian, Yang and Zhu).
(Mousavi, Abdollahinejad and Corizzo) propose E-CaTCH, a multimodal learning framework designed to detect misinformation on social media while addressing two key challenges: inconsistencies across modalities (text and images) and severe class imbalance in real-world datasets. Unlike traditional approaches that treat posts independently, E-CaTCH models misinformation at the event level, where related posts are grouped into "pseudo-events" based on textual similarity and temporal proximity. Each event is processed separately to better capture evolving narratives. Within each event, the model extracts features using pretrained encoders (BERT for text and ResNet for images). These representations are refined using self-attention and aligned through cross-modal attention, allowing better interaction between modalities. A soft gating mechanism then fuses the modalities into a unified representation. To capture how misinformation evolves over time, the framework applies a trend-aware LSTM over event sequences. To address class imbalance, it integrates adaptive class weighting and temporal consistency regularisation, which stabilises learning and improves detection of rare classes. Experimental results show that E-CaTCH achieves strong performance (including state-of-the-art accuracy on benchmark datasets) and generalises well across different misinformation scenarios (Mousavi, Abdollahinejad and Corizzo).

2.6. Sampling strategies for rare events

(Harada) reviews enhanced sampling methods designed to overcome the limitations of conventional molecular dynamics simulations in capturing rare biological events, particularly protein conformational transitions that occur infrequently but are crucial for biological function. Since these transitions happen on long timescales that standard simulations cannot efficiently reach, the study focuses on techniques that increase the probability of observing such events through improved sampling strategies. The author discusses several approaches, including Parallel Cascade Selection Molecular Dynamics (PaCS-MD), fluctuation flooding, taboo search, outlier flooding, structural dissimilarity sampling, and self-avoiding conformational sampling. These methods generally rely on a cycle of short simulations followed by the selection of promising intermediate states, which are then used to restart new simulations. This iterative resampling process gradually guides the system toward rare transitions that would otherwise be difficult to observe. Overall, the study demonstrates that enhanced sampling strategies significantly improve the ability to model rare events in protein systems, making computational predictions more efficient and biologically meaningful (Harada).
(Webber, Plotkin and O'Neill) address the challenge of simulating rare and extreme events in complex physical systems, with a focus on mesoscale weather phenomena such as tropical cyclones, squall lines, and floods. These events are difficult to capture using standard simulation methods because they occur infrequently and lie in the extreme tails of probability distributions. To

overcome this limitation, the authors introduce a rare event sampling method called Quantile Diffusion Monte Carlo (quantile DMC), which is designed to efficiently generate extreme realisations of stochastic processes. The method works by preferentially sampling high-impact, low-probability outcomes, allowing researchers to obtain statistically meaningful estimates of extreme weather behaviour with reduced computational cost. The paper demonstrates the effectiveness of quantile DMC by applying it to historical hurricanes, showing that it can successfully reproduce extreme scenarios that traditional simulation approaches would rarely capture. Overall, the study highlights how advanced sampling strategies can improve the efficiency and accuracy of modelling rare environmental events, particularly in climate and weather systems (Webber, Plotkin and O'Neill).
(W. Zhang, H. Wang and C. Hartmann) in his article titled Applications of the Cross-Entropy Method to Importance Sampling and Optimal Control of Diffusions presents a computational framework for efficiently estimating rare event probabilities in diffusion processes. The key idea is the use of the cross-entropy method, which constructs an optimised "tilted" probability distribution that biases sampling toward rare but important events while maintaining statistical correctness. This reduces the number of samples needed compared to standard Monte Carlo simulation. The method is also extended to optimal control problems, showing how rare-event simulation techniques can be reformulated as optimisation problems over probability distributions. Overall, the work demonstrates how adaptive importance sampling can significantly improve efficiency in modelling low-probability, high-impact events in complex stochastic systems.
(Plattner, Doll and Dupuis) in his article titled An Infinite Swapping Approach to the Rare-Event Sampling Problem introduces a computational technique designed to improve the efficiency of sampling rare events in Monte Carlo simulations. Traditional sampling methods struggle with rare events because the relevant configurations are poorly explored due to energy barriers or sparse probability regions. The authors address this by proposing an "infinite swapping" symmetrisation strategy, which transforms the original probability distribution into a more connected and smoothly navigable form. This makes it easier for sampling algorithms to explore otherwise inaccessible regions of the state space. The method is demonstrated on molecular systems (such as Lennard-Jones clusters), showing improved ability to capture rare transitions and enhance statistical convergence. Overall, the work contributes a novel theoretical and practical framework for improving rare-event simulation in statistical physics and computational chemistry.

2.7. Manual vs semiautomatic annotation

(Naziroglu, Puylaert and Tielbeek) in his article titled Semi-automatic bowel wall thickness measurements on MR enterorrhaphy in patients with Crohn's disease directly compares manual annotation by radiologists with a semi-automatic segmentation method for medical image analysis. In this work, human experts manually delineate regions of Crohn's disease on MRI scans and perform thickness measurements, while the proposed system uses a semi-automatic active contour segmentation approach that is still supervised by observers but reduces manual effort. The results show that the semi-automatic method produces higher reproducibility, lower variance, and better interobserver agreement compared to fully manual annotation. This positions the work squarely in the category of semi-automatic annotation systems evaluated against manual ground truth, rather than fully automated segmentation or purely manual labelling studies.
(Ghosh and Jiang) in their article titled Artificial intelligence framework to identify spatial patterns of cancer immunotherapy outcomes focuses on an AI-based framework for analysing multiplexed proteomics imaging to identify spatial patterns associated with immunotherapy response or resistance. Unlike classical annotation studies, it explicitly avoids human-driven steps such as manual or semi-automatic segmentation and cell labelling and instead uses a fully automated region-based deep learning approach to extract features directly from images. So, while it is related to medical imaging and spatial analysis, it is not a manual vs semi-automatic annotation comparison

paper. Instead, it belongs more to fully automatic AI-driven image analysis / weakly supervised spatial feature learning, where the goal is to eliminate the need for manual or semi-automatic annotation altogether.

(Weizman, Hoch and Ben Sira) propose a semi-automatic method for the classification and segmentation of Plexiform Neurofibromas (PN), which are peripheral nerve sheath tumours associated with Neurofibromatosis type-1. The study addresses a key clinical challenge: manual MRI-based tumour segmentation is highly time-consuming, error-prone, and requires extensive expert effort for accurate volume estimation, which is essential for treatment decisions. The proposed approach reduces this burden by requiring minimal user interaction. A clinician manually outlines the tumour region on a single MRI slice, and the algorithm automatically propagates the segmentation across the remaining slices using tumour connectivity and image processing techniques. This semi-automatic pipeline enables efficient detection and quantification of tumour volume across the full scan. Experimental evaluation on multiple patient datasets shows that the method achieves results reasonably close to expert manual segmentation, with an average volume overlap difference of about 25%. Although not fully automated, the method significantly reduces annotation time, from over an hour for full manual segmentation to approximately 12 minutes, while maintaining clinically useful accuracy. Overall, the study demonstrates that semi-automatic segmentation can effectively balance efficiency and reliability, making it a practical tool for supporting clinical workflow in medical image analysis.

2.8. Cross modal labelling consistency

(Zhu, Hu and Lu) propose a LiDAR-free radar–visual auto-labelling framework designed to improve the quality of 3D annotation in autonomous driving scenes by enforcing cross-modal consistency between radar and vision data. The method addresses key limitations of vision foundation models, which can generate pseudo-3D point clouds that suffer from scale ambiguity, structural noise, and temporal inconsistency. Instead of relying on expensive LiDAR sensors, the framework integrates millimetre-wave radar trajectories with vision-based 2D instance masks and lifted pseudo-point clouds. The pipeline first constructs object-centric temporal sequences by associating radar points, visual masks, and pseudo-3D geometry. It then applies an uncertainty-aware pose fusion module that combines motion information from radar with structural cues from vision, guided by road geometry priors to ensure physically plausible scene understanding. Finally, the pseudo-point clouds are refined in a canonical space by extracting stable semantic landmarks from temporally consistent masks and propagating corrections across time. Experiments on a real-world roadside dataset show that the method achieves 49.1% BEV IoU and 43.0% 3D IoU, outperforming a radar–camera baseline by 5.5 and 5.9 points respectively. Overall, the study demonstrates that enforcing cross-modal spatiotemporal consistency significantly improves automated 3D labelling quality, making it a promising direction for scalable dataset annotation in autonomous driving.

(Yang, Wei and Zhu) propose a semi-supervised image captioning method that improves learning from limited labelled data by enforcing cross-modal prediction and relation consistency between images and generated text. The main challenge addressed in the study is that image captioning typically requires large amounts of manually annotated image–sentence pairs, which is expensive and difficult to scale. To overcome this, the authors introduce a framework that leverages both labelled and unlabelled images more effectively. Their method uses two key ideas: prediction consistency, where the model uses image predictions as soft supervision signals to guide sentence generation, and relation consistency, which ensures that semantic relationships between objects in images are preserved in the generated captions. Both images and sentences are mapped into a shared semantic space to reduce the modality gap and improve alignment. By combining these two consistency constraints, the model can better learn from unlabelled data without relying heavily on manual annotation or traditional pseudo-labelling techniques. Experiments on benchmark datasets

(e.g., MS-COCO) show that the approach significantly improves caption quality, achieving notable gains (around 6% in CIDEr-D score) compared to existing methods. Overall, the study demonstrates that cross-modal consistency mechanisms can effectively reduce annotation dependency and improve semi-supervised vision–language learning.
(Pu, Su and Hu) in their paper introduces NOTO, a framework designed to improve cross-modal retrieval when training data contains both label noise and open-set samples (unknown classes not included in training). These issues typically degrade performance because they corrupt the shared semantic space between modalities such as images and text. To address this, the authors propose a Robust Evidential Learning (REL) module that estimates uncertainty using Dirichlet evidence, allowing the model to distinguish clean, noisy, and open-set instances. The system then applies different learning strategies depending on the reliability of each sample. In addition, the Adaptive Noise-aware Contrast (ANC) module improves cross-modal alignment by selectively using reliable sample pairs and reinforcing consistency through contrastive learning while reducing the impact of noisy labels. Overall, the method improves robustness and retrieval accuracy across multiple benchmark datasets, outperforming existing cross-modal retrieval approaches in noisy and open-set conditions.

2.9. Weak supervision and noisy labels

(Takasan and Iiyama) addresses the problem of estimating fishing ground locations when training data is limited and labels are weak, incomplete, or noisy. Improving fishing ground estimation with weak supervision and meta-learning. The study focuses on the challenge of building accurate machine learning models for fishing ground prediction, where obtaining fully labelled catch data is expensive and only partially available. To overcome this, the authors propose a framework that combines weak supervision with meta-learning. Weak supervision is used to incorporate additional but imperfect data sources, such as sea surface temperature patterns and trajectory data, which provide indirect and noisy signals about fishing locations. While this increases available training data, it also introduces uncertainty and label noise. To mitigate this issue, the authors introduce a meta-learning strategy that helps the model adapt more effectively to noisy supervision by dynamically adjusting learning behaviour during training. The approach includes a two-stage pipeline: pre-training on large-scale trajectory data followed by fine-tuning on more reliable catch data, with meta-learning guiding the optimisation process. Experimental results demonstrate that the proposed method significantly improves predictive performance, achieving a 64% increase in F1-score compared to baseline methods that rely only on limited labelled data. The study concludes that combining weak supervision with meta-learning is an effective strategy for handling noisy and incomplete annotations in real-world data settings.
(Mwubahimana, Jianguo and Miao) addresses the problem of learning high-resolution land cover maps when training data is limited, coarse, or noisy under weak supervision.C2FNet: Cross-Probabilistic Weak Supervision Learning for High-Resolution Land Cover Enhancement. The study focuses on the challenge of generating accurate high-resolution land cover (HRLC) maps from low-quality or coarse annotations, which often contain label noise and resolution mismatches. These limitations make it difficult for standard supervised learning methods to produce reliable fine-grained spatial outputs. To solve this, the authors propose C2FNet (Coarse-to-Fine Network), a weakly supervised framework that refines coarse labels into high-resolution predictions. The model includes multiple components: edge-aware feature extraction to preserve spatial details, spatial consistency–based confidence estimation to identify reliable regions, and a hybrid loss function combining cross-entropy and similarity-based constraints to reduce noise impact. The system is trained using a coarse-to-fine strategy, where noisy or low-resolution labels guide initial learning and progressively refined representations improve spatial accuracy. Experiments across multiple datasets show that the proposed method achieves state-of-the-art performance, significantly

improving overall accuracy and robustness compared to existing weak supervision approaches. In summary, the paper demonstrates that carefully designed weak supervision strategies can effectively overcome label noise and resolution limitations in large-scale remote sensing applications.

(Nashaat, Ghosh and Miller) addresses the challenge of expensive and time-consuming manual data annotation in machine learning, where weak supervision is used to generate large-scale training datasets but often introduces noisy and unreliable labels. Asterisk: Generating Large Training Datasets with Automatic Active Supervision. The study focuses on improving the quality of weakly supervised datasets by proposing Asterisk, an end-to-end framework that automatically generates and refines training labels. Instead of relying solely on manual annotation, the system first uses heuristic-based rules to assign initial labels, enabling large-scale dataset creation at low cost. However, since these heuristics can introduce noise, the framework incorporates a data-driven active learning strategy to iteratively identify and refine the most informative and uncertain samples. In addition, Asterisk models the reliability of different labelling heuristics and uses this information to guide the learning process, improving the overall consistency of the generated labels. Through repeated refinement, the system enhances both label accuracy and model performance while significantly reducing the need for human annotation. Experimental evaluations on multiple large-scale datasets demonstrate that Asterisk achieves higher labelling quality and improved classification performance compared to existing weak supervision methods, including datasets with millions of records, highlighting its effectiveness in scalable and cost-efficient data annotation.

3. Data Quality, Validation, and Provenance in Multimodal Datasets

3.1. Completeness, consistency, accuracy, timeliness

(Lasim, Ansah and Apaak) examine maternal and child health (MCH) data quality in healthcare facilities in the Cape Coast Metropolis, Ghana, focusing on how well routine health information systems reflect real clinical activity. The study evaluates data quality using four key dimensions: accuracy, completeness, timeliness, and consistency. The authors conducted a facility-based cross-sectional review of health records across 13 healthcare facilities. They compared data from multiple sources, including facility registers, monthly reporting forms, and the District Health Information System (DHIS2). Eight MCH indicators (such as immunisation and maternal service records) were used to assess accuracy and consistency, while completeness and timeliness were evaluated using monthly reporting submissions. Results showed generally high data quality across systems. Accuracy between data sources was close to or slightly above 100%, indicating strong agreement between registers, forms, and DHIS2 records. Completeness ranged from about 91% to 95% depending on the source, while timeliness averaged around 87% for monthly reporting. Consistency was also high at approximately 93%. Overall, the study concludes that MCH data in these facilities is largely reliable and reflective of actual service delivery, although improvements are still needed, particularly in reporting timeliness and uniform data recording practices.

(Ghalavand, Shirshahi and Rahimi) present a systematic review that identifies and consolidates the key data quality dimensions used in health information systems. The study focuses on clarifying how data quality is defined and evaluated across existing literature, given the lack of standardisation in the field. The authors conducted a structured literature review using major academic databases (Web of Knowledge, Scopus, PubMed, ScienceDirect, Emerald, and Google Scholar). From 760 initial papers, 58 studies were selected for detailed analysis after removing duplicates and applying inclusion criteria. The review synthesises and categorises data quality dimensions used across health information system research. The findings show that 14 core data quality dimensions are commonly used, including accuracy, consistency, completeness, timeliness, reliability, accessibility,

objectivity, relevance, understandability, navigation, reputation, efficiency, and value-added. Among these, accuracy, completeness, and timeliness are the most frequently used and widely accepted core indicators. The study concludes that although many dimensions exist, there is still no universal standard for measuring data quality, and different studies apply varying definitions and evaluation approaches. This inconsistency highlights the need for a more unified framework for assessing data quality in health information systems.

(Phinias, Muhanga and Malago) investigate variability in healthcare data quality across regional referral hospitals in Tanzania, focusing on how healthcare workers perceive four key data quality dimensions: accuracy, completeness, timeliness, and consistency. The study highlights that data quality is essential for effective healthcare decision-making, but its perceived quality can differ significantly across institutions. The authors conducted a cross-sectional survey involving 336 healthcare workers from eight regional referral hospitals. Participants completed a validated questionnaire assessing perceptions of the four data quality dimensions. Statistical analyses, including ANOVA and post-hoc tests, were used to examine differences across hospitals and dimensions. The results show significant variability in perceived data quality across hospitals. All four dimensions, accuracy, completeness, timeliness, and consistency, showed statistically significant differences between facilities, with medium-to-large effect sizes. Some hospitals consistently scored higher across all dimensions, particularly in completeness, while others showed lower performance. The study concludes that data quality is not uniform across healthcare facilities and emphasises the need for integrating objective measurement tools (such as audit logs and error rates) alongside perception-based assessments. It also recommends using structured frameworks (e.g., Donabedian model) to better understand the institutional factors affecting data quality and to improve healthcare data reliability.

Modality-specific validation

(Chen, Jia and Zhou) propose a hierarchical spatiotemporal graph learning framework for risk prediction in complex systems where multiple risk factors co-occur and evolve over time. The study focuses on improving prediction accuracy in domains such as medical diagnosis and system safety by addressing the limitation of traditional models that treat spatial relationships and temporal dynamics separately. The authors design a model that integrates both spatial co-occurrence patterns and temporal progression patterns. First, a dual-matrix graph construction mechanism captures spatial risk correlations (based on co-occurrence frequency) and temporal transitions (based on progression probabilities). Second, an adaptive subgraph extraction module builds system-specific representations that preserve both localised risk clusters and long-term temporal pathways. Third, a dual-channel graph convolutional network fuses spatial and temporal features while maintaining modality-specific information. Experimental results across multiple domains show that the proposed model significantly outperforms conventional single-modality approaches, demonstrating strong effectiveness in handling complex multi-risk interactions and long-term dependencies. The study concludes that combining spatial and temporal representations in a structured graph framework improves robustness and generalizability in cross-domain risk prediction tasks.

(Rajaraman and Antani) propose a modality-specific deep learning framework to improve tuberculosis (TB) detection in chest radiographs by leveraging knowledge transfer across multiple imaging datasets. The study trains convolutional neural networks (CNNs) on several large-scale chest X-ray datasets to learn modality-specific feature representations before transferring the learned knowledge to a target TB classification task. This approach enables the model to capture modality-dependent characteristics while maintaining generalisation across datasets. The predictions from multiple modality-specific models are further combined using ensemble learning strategies to improve robustness and reduce prediction variance. Experimental results demonstrate that the stacked ensemble achieves high classification performance (accuracy ≈ 0.941, AUC ≈ 0.995), outperforming individual models and improving stability under cross-validation. The study

concludes that modality-specific learning combined with ensemble validation enhances both predictive accuracy and generalisation in medical imaging applications.
(Zhang, Lu and Chen) propose a Multimodality Distillation and Fusion Network (ModiF) for age-related macular degeneration (AMD) classification using multiple retinal imaging modalities, including OCT, OCTA, and CFP. The study addresses the challenge that existing approaches often fail to fully exploit complementary information across modalities due to heterogeneous data distributions and complex lesion patterns. To overcome this, the model introduces modality-specific feature extractors that learn hierarchical semantic representations from each imaging modality independently. These representations are then enhanced using a cross-modal soft supervision strategy, where modality-specific predictions are adaptively combined (voting-based) and used as soft targets for knowledge distillation during training. The framework further includes a principal-modality-guided distillation mechanism, which facilitates controlled cross-modal knowledge transfer while preserving critical diagnostic features. In addition, a confidence-aware fusion module dynamically adjusts the contribution of each modality based on prediction entropy, ensuring more reliable integration of multimodal information. Experimental results across one internal dataset and two public datasets demonstrate that the proposed method consistently outperforms state-of-the-art approaches, showing strong generalisation ability and improved robustness in AMD classification tasks.

3.2. Cross-modal consistency validation

(Mazhar, Zada and Aldhayan) propose a hybrid deep learning framework for detecting cyberbullying in short-form video content, addressing the challenges posed by multimodal abuse signals on platforms such as Instagram Reels, TikTok, and YouTube Shorts. The study highlights that cyberbullying detection is complex because harmful content is expressed through multiple synchronised modalities, including visual cues, audio patterns, textual captions, and spoken language. To handle this, the authors integrate Convolutional Neural Networks (CNNs) for spatial visual feature extraction, Bidirectional Long Short-Term Memory (BiLSTM) networks for modelling temporal acoustic patterns, and a Transformer-based encoder for textual analysis. A key contribution is the semantic-consistency validation layer, which enforces alignment across modalities using attention-based similarity constraints, penalising inconsistent cross-modal signals to improve classification reliability. Experimental results on the CAVD and SocialVidMix datasets demonstrate strong performance, achieving 91.6% accuracy and an F1-score of 91.3%, with robust generalisation across multiple platforms. The authors conclude that enforcing cross-modal consistency significantly improves the robustness and scalability of multimodal cyberbullying detection systems.
(Tao, Zhu and Cao) investigate fine-grained cross-modal semantic consistency in multimodal learning, focusing on image–text alignment challenges in complex datasets such as natural conservation imagery. The study highlights that cross-modal representation learning often suffers from semantic ambiguity and encoder-induced inconsistencies, which can reduce the reliability of similarity matching across modalities. To address this issue, the authors propose a multi-task learning framework that integrates cross-modal retrieval and generation tasks to improve semantic alignment. A residual attention network is introduced to stabilise encoder updates during momentum-based learning, reducing representation drift and improving consistency between modalities. Additionally, a momentum encoding mechanism is used to bridge information across modalities and enhance mutual information in the shared embedding space. This approach helps mitigate contextual ambiguity in natural language descriptions while preserving invariant visual features. Experimental results demonstrate that the proposed method improves cross-modal alignment performance by up to 8% on benchmark tasks, showing stronger retrieval accuracy and more robust semantic representation compared to baseline models. The study concludes that

enforcing cross-modal consistency through multi-task momentum encoding significantly enhances the quality and robustness of multimodal representations.
(Hou, Chen and Shang) propose a cross-modal consistency-driven, quality-aware dynamic fusion framework for emergency damage assessment using multimodal inputs such as visual, textual, and audio data. The study addresses key challenges in multimodal learning, including inconsistent data quality, missing modalities, and noise, which often reduce the reliability of fusion-based prediction systems. The proposed method first maps each modality into a shared semantic representation space and applies cross-modal reconstruction learning to evaluate how well different modalities align with each other. The reconstruction error is then used as a quality indicator for each modality. A dynamic gating mechanism combines these quality signals with confidence measures to assign adaptive weights to each modality during fusion. Experimental results under different conditions (complete, missing, and degraded modalities) demonstrate that the framework improves accuracy, Macro-F1 score, and robustness, while reducing the impact of low-quality or unreliable modalities. Overall, the study shows that explicitly modelling cross-modal consistency and modality quality leads to more stable and reliable multimodal decision-making in emergency scenarios.

3.3. Metadata standards

(Vangay, Burgin and Johnston) emphasise the importance of metadata standards for ensuring that microbiome data are well described, interpretable, and reusable. The authors explain that metadata provide essential contextual information about biological samples, enabling accurate analysis, data sharing, and reproducibility. Although several community-driven metadata standards have been developed, their adoption across the microbiome research community remains inconsistent. Based on discussions from the National Microbiome Data Collaborative (NMDC) workshop, the study identifies challenges to metadata standard implementation and proposes recommendations to improve standardisation, interoperability, and collaboration among researchers, institutions, and funding agencies.
(Vangay, Burgin and Johnston) reviewed the current state of metadata standards in microbiome research and highlighted the importance of standardised metadata for improving data interpretation, sharing, and reproducibility. The authors noted that although several community-driven metadata standards have been developed, their adoption remains inconsistent across research institutions and scientific domains. Based on discussions from the National Microbiome Data Collaborative (NMDC) workshop, the paper identifies major barriers to metadata standardisation, including inconsistent implementation, limited awareness, and lack of coordination among stakeholders. The authors recommend greater collaboration among researchers, funding agencies, and data repositories to establish common metadata standards and best practices that support data interoperability and reuse.
(Branton, Worcester and Lafond) summarised the outcomes of a technical symposium and workshop focused on the role of metadata standards in achieving data interoperability. The proceedings emphasise that standardised metadata is essential for enabling data discovery, exchange, accessibility, and integration across different information systems. The symposium covered topics such as metadata relevance, interoperability, authority control, and open-access metadata systems. The workshop further discussed key issues including the definition of metadata units, the selection of appropriate metadata attributes, and the use of classification and taxonomic systems to ensure consistent and interoperable data management.
(Rokem, Mandava and Cristea) reviewed how open-source software (OSS) development practices can improve the creation and maintenance of data and metadata standards across scientific disciplines. The authors argue that interoperable metadata standards are essential for integrating heterogeneous datasets, ensuring reproducibility, and supporting AI- and machine learning-driven research. Drawing examples from neuroscience, astronomy, earth science, and high-energy

physics, they show that OSS practices such as version control, community collaboration, and consensus-driven development make metadata standards more adaptable, transparent, and sustainable. The review also discusses challenges, including balancing flexibility with stability and ensuring long-term maintenance, while recommending greater institutional support for community-driven, open-source metadata standards.

3.4. Lineage tracking across data pipelines

(Jacques-Silva, Kalyvianaki and Cohn-Gordon) proposed the Unified Lineage System (ULS), an end-to-end data lineage framework designed to track data provenance across large-scale, heterogeneous data pipelines. The system introduces a generalised data model for representing data flows across different storage assets, a lineage model that supports navigation across multiple levels of granularity, and techniques for reconstructing incomplete lineage traces. It also incorporates graph-processing methods to balance precision and recall according to application requirements. Deployed at Meta, ULS manages billions of lineage nodes and edges, supporting applications such as operational monitoring, capacity management, and privacy compliance. The study demonstrates that scalable lineage tracking significantly improves data transparency, traceability, and governance in complex data ecosystems.
(Girten) presents a comprehensive guide to building and managing modern data pipelines using the Databricks Lakehouse platform. The book discusses the complete data engineering lifecycle, including real-time data ingestion, transformation, orchestration, monitoring, governance, and deployment. A key contribution is its explanation of Unity Catalog, which enables data lineage tracking by visualising how data flows across datasets, transformations, and pipelines. The book also covers data validation, automated workflow management, CI/CD integration, and data governance practices, demonstrating how lineage tracking improves transparency, debugging, compliance, and trust in large-scale data engineering environments.

3.5. Dataset cards and model cards

(Singh, Lodwal and Malwat) introduce a dataset for automated model card generation using question–answer pairs extracted from machine learning research papers. Model cards are structured documents that describe key information about ML models, such as training data, architecture, intended use, limitations, and potential biases. Creating them manually is time-consuming and inconsistent. To address this, the authors build a dataset of 500 question–answer pairs across 25 ML models, where each pair captures important model details extracted from research papers. The dataset is intended to train or evaluate language models in generating accurate and structured model cards automatically. The paper also tests LLMs (e.g., ChatGPT-3.5 and LLaMA) and finds that current models struggle with producing fully accurate and reliable model documentation, highlighting a gap in factual understanding and structured generation. The authors conclude that their dataset can help improve automated documentation systems and reduce the manual effort required in ML model reporting.
(Ango, Masih and Reddy) present a machine learning approach for detecting fraudulent credit card transactions using the Kaggle Credit Card dataset. The dataset contains highly imbalanced transaction records labelled as either fraud or non-fraud, reflecting real-world financial data challenges. The authors apply the XGBoost classifier, an ensemble learning method known for its strong predictive performance and efficiency. Through feature engineering and hyperparameter tuning, the model is optimised to improve fraud detection accuracy while minimising false positives. The results demonstrate strong performance, with the model achieving approximately 98.6% accuracy, along with high precision, recall, and F1-score values. The study concludes that XGBoost

is highly effective for improving fraud detection systems in banking and financial security applications.
(Alam, Shaukat and Hameed) investigate credit card default prediction using multiple imbalanced financial datasets, focusing on improving classifier performance under data imbalance conditions. The study highlights that traditional statistical methods are insufficient for handling imbalanced credit risk data, where one class dominates the dataset. To address this, the authors apply data preprocessing techniques such as normalisation and resampling methods (oversampling and undersampling) to balance the dataset. They then evaluate multiple machine learning models, including ensemble-based approaches, to determine their effectiveness in predicting credit default. The results show that handling class imbalance significantly improves model performance. Accuracy increases from around 65–70% on imbalanced data to nearly 85–89% on balanced datasets. Among the tested methods, Gradient Boosted Decision Trees combined with SMOTE-based oversampling perform the best. The study concludes that resampling techniques and ensemble models are highly effective for improving credit risk prediction systems.

3.6. Industrial best practices

(Cardin, Ounnar and Thomas) examine best practices in future industrial systems through the work of the IMS2 French research group, focusing on intelligent manufacturing and service systems. The study highlights key industrial priorities including agility, technological innovation, sustainability, and knowledge dissemination within Industry 4.0 environments. The authors synthesise multiple research contributions to identify effective strategies for developing smart industrial systems. They also evaluate existing approaches and propose a future roadmap for industrial development toward 2030, emphasising the integration of advanced technologies such as automation, data-driven manufacturing, and intelligent coordination systems. The paper concludes that adopting structured best practices is essential for improving efficiency, adaptability, and sustainability in modern industrial ecosystems.
(Najjar and Abu-Shamleh) evaluate the performance of a large-scale thermal power plant using internationally recognised industrial best practices. The study analyses key operational metrics, including availability, reliability, capacity factor, and thermal efficiency, over an eight-year period (2010–2017). The authors compare the plant's performance against global benchmark standards to assess operational efficiency and identify gaps. The results show that while the plant performs reasonably well in terms of capacity factor, it falls short in areas such as availability, reliability, and thermal efficiency when compared to industry best-practice targets. The study concludes that adherence to industrial best practices is essential for improving efficiency, reducing power losses, and enhancing the overall performance of large-scale energy systems.
(Zhao, Blackburn and McKinley) present a study on improving software system performance by systematically combining research methodologies with industrial best practices. The paper focuses on bridging the gap between academic research and real-world production systems, where deployment constraints and complexity often limit direct application of research results. The authors describe their experience translating a high-performance garbage collection algorithm into a hyperscale production environment. By integrating research-driven techniques with industrial engineering practices, they achieved significant performance improvements, including reducing garbage collection costs by nearly 50% and outperforming existing state-of-the-art systems by over 10%.The study also highlights key lessons from this process, emphasising the importance of aligning research metrics with production requirements, using realistic workload patterns, and addressing real-world system constraints. The paper concludes that combining systematic research practices with industrial best practices can significantly enhance both performance and real-world applicability.

4. Confidentiality, Privacy, and Secure Data Handling

4.1. Data confidentiality in industrial AI

(Nguyen, Ding and Pathirana) discuss the role of federated learning in Industrial Internet of Things (IIoT) systems to address challenges related to data confidentiality and large-scale distributed AI training. The study highlights that traditional AI methods rely on centralised data collection, which is often not feasible in industrial environments due to privacy, security, and scalability concerns. To overcome this, the authors propose federated learning, where AI models are trained locally on industrial devices and only model updates are shared, rather than raw data. This approach helps preserve sensitive industrial information while still enabling collaborative machine learning across distributed systems. The paper concludes that federated learning is a promising solution for maintaining data confidentiality in industrial AI systems, while supporting scalable and efficient intelligent industrial applications.

(Arpaci) investigates the social sustainability of AI chatbots by examining how cybersecurity factors, particularly data confidentiality and privacy, influence user adoption and trust. The study builds a model based on Protection Motivation Theory and evaluates it using structural equation modelling and artificial neural networks. Using data from 1,741 participants, the research finds that confidentiality and privacy concerns are key predictors of sustainable chatbot use. The results show that cybersecurity significantly impacts both user behaviour and the long-term adoption of AI systems, explaining a large portion of variance in sustainability outcomes. The study concludes that ensuring data confidentiality is essential for the safe and sustainable deployment of AI chatbot technologies.

(Zolanvari) examines the challenges of applying artificial intelligence to the security of Industrial Internet of Things (IIoT) and industrial control systems. The study highlights that industrial environments face significant barriers due to data confidentiality, which prevents organisations from sharing real operational data for research and model training. The dissertation identifies key limitations in industrial AI systems, including data scarcity, black-box model behaviour, and high computational cost. It also presents approaches to address these challenges, such as building realistic industrial testbeds for data collection, developing explainable AI methods, and using distributed AI systems to improve efficiency and security. The work concludes that overcoming data confidentiality constraints is essential for enabling reliable and scalable AI-based security solutions in industrial systems.

(Le Roux, Khaksar and Sepehri) propose a hybrid AI framework to address data scarcity in open-pit mining crack detection, where real-world datasets are limited due to safety risks, high costs, and data confidentiality constraints. The study combines a Unreal Engine 5 (UE5) simulated environment with StyleGAN2-ADA to generate realistic synthetic images of mining cracks. These generated datasets are used to train a YOLOv11 object detection model, which is evaluated on real-world images. Results show significant performance improvements, demonstrating that synthetic data generation can effectively reduce dependence on sensitive industrial datasets while maintaining strong real-world generalisation.

4.2. Role-based access and data partitioning

(Zhong, Lentz and Narodytska) propose HONEYBEE, a dynamic partitioning framework designed to improve role-based access control (RBAC) in vector databases. The work addresses a key trade-off in secure data systems: per-user indexing improves query speed but increases memory usage, while shared indexing reduces memory use but slows down queries due to filtering overhead. HONEYBEE leverages the structure of RBAC policies, where users are grouped into roles with shared permissions, to create overlapping data partitions. This allows selective replication of vectors

across partitions to balance query efficiency, memory consumption, and retrieval accuracy. The framework uses analytical performance models and optimisation techniques to determine optimal partitioning strategies. Experimental results show that HONEYBEE significantly improves query latency while maintaining low memory overhead, making it a practical solution for secure and efficient vector database management in enterprise environments.
(Arshad, Felemban and Pervaiz) propose a privacy-preserving framework for managing access-controlled graph data using role-based access control (RBAC). The study focuses on ensuring that only authorised users can access specific parts of graph data based on their assigned roles, while still protecting sensitive information from unauthorised disclosure. The authors introduce a graph partitioning approach (k-anonymous bi-objective graph partitioning, k-BGP) that balances two competing goals: privacy protection and authorised access privileges. To achieve this, the system partitions graph data into anonymised views based on roles, while applying heuristics to handle the trade-off between usability and privacy. The framework is evaluated for its resistance to re-identification attacks and demonstrates how role-based structures can be used to enforce secure data sharing.

4.3. Anonymisation and pseudonymisation

(Raso, Loreti and Ravaziol) propose a flexible framework for anonymisation and pseudonymisation of healthcare data using FHIR (Fast Healthcare Interoperability Resources) to support the secondary use of Electronic Health Records in research and healthcare analytics. While FHIR enables efficient data exchange between systems, privacy regulations limit its use because patient data can contain direct or indirect identifiers. To address this, the study introduces FHIR-DIET, a configurable de-identification system that applies anonymisation and pseudonymisation techniques to medical data. The framework allows customizable privacy rules so that legal and technical teams can collaboratively define how sensitive information is processed. The system is designed to maintain data utility while ensuring compliance with privacy regulations. Performance evaluation shows that it can process large volumes of patient data efficiently, making it suitable for real-world healthcare environments.
(Alicea) in their article titled, Enhancing Social Media User Privacy: Applying Machine Learning-Based Anonymisation and Pseudonymisation Techniques proposes a machine-learning-based privacy framework to protect sensitive user data in social media and e-commerce systems while preserving data utility for analysis. The study combines Named Entity Recognition (NER) to detect personal identifiable information, cryptographic hashing (SHA-256) for pseudonymisation, and k-anonymity techniques to reduce re-identification risk. Evaluated on large datasets, the approach achieves around 88–90% risk reduction with only a small drop in model performance, showing that strong privacy protection can be achieved without significantly affecting analytical accuracy.
(Dimopoulou, Symvoulidis and Koutsoukos) in his article titled, Mobile Anonymisation and Pseudonymisation of Structured Health Data for Research proposes a privacy-preserving framework for healthcare data sharing using anonymisation and pseudonymisation techniques within the Fast Healthcare Interoperability Resources (FHIR) standard. The study presents a mobile library that enables secure transformation of patient data before it is shared with research centres, helping reduce the risk of identity disclosure while maintaining data usability. Two case studies demonstrate how the system can effectively support both anonymisation and pseudonymisation in real-world healthcare data environments.

4.4. Data minimisation

(T. Li, Z. Chen and Z. Zhang) in his article titled, Predicting Stress–Strain Curve with Confidence: Balance Between Data Minimisation and Uncertainty Quantification by a Dual Bayesian Model presents a machine learning framework that balances data minimisation with uncertainty-aware

prediction in polymer engineering. The study uses a small, carefully designed dataset (Taguchi array) to reduce the amount of required training data while still maintaining high predictive performance for stress–strain curve estimation. A dual Bayesian neural network is applied to capture both aleatoric and epistemic uncertainty, allowing the model to provide reliable predictions with confidence intervals. Overall, the work shows that accurate ML modelling is possible even with limited data when uncertainty quantification is properly integrated.

(Bodhe, Amyeen and Pomeranz) proposed a fail-data minimisation method using an N-cover algorithm to reduce unnecessary diagnostic data in large-scale integrated circuit testing. The approach focuses on eliminating non-contributing failing test information while preserving diagnostic accuracy. By minimising redundant fail data, the method significantly improves the efficiency and speed of the diagnosis process without reducing reliability. Experimental results show that the technique achieves around 40% reduction in fail data while maintaining about 95% diagnostic accuracy, leading to faster and more efficient hardware failure analysis.

(Pu, Chen and Pan) proposed a facial data minimisation framework that acts as a privacy-preserving filter for face recognition systems. The method introduces a Privacy Minimisation Transformation (PMT) that processes facial images or features using a shallow model to produce obfuscated representations. These transformed outputs preserve performance for authorised recognition systems while reducing effectiveness for unauthorised models and preventing privacy leakage. The study also introduces enhanced perturbation and multi-restriction mechanisms to improve robustness and scalability. Overall, the approach helps minimise facial data exposure while maintaining high recognition accuracy and reducing risks such as function creep and biometric misuse.

4.5. Federated data access

(Bienzeisler, Kombeiz and Ehrentreich) proposed a federated data access authorisation system that enables secure, large-scale sharing of electronic health records while keeping data locally controlled. The framework is designed to balance data accessibility with privacy and governance requirements by allowing remote queries without transferring raw patient data. It supports recurring and periodic queries, enabling near real-time access for authorised users across distributed healthcare institutions. The implementation in the German National Emergency Department Data Registry demonstrated scalability and efficiency, successfully managing millions of records and handling increasing query loads while maintaining controlled access and regulatory compliance.

(Miyamoto and Kasuga) proposed an Enterprise Data Science Platform (EDSP) to solve the problem of data silos in organisations where different departments use different analytics tools. Traditional systems often require copying datasets into multiple platforms, which leads to duplication, inconsistencies, and higher management costs. The proposed EDSP uses a federated data access architecture based on a “write-once, read-anywhere” model. Instead of moving or duplicating data, datasets remain in a centralised data store and are accessed directly by multiple query engines through a unified access layer. The system is built with four layers: data preparation, data store, access interface, and query engines. The study shows that EDSP reduces operational steps by 33–44% compared to conventional data migration approaches and supports interoperability across different analytical platforms. Although query latency can increase slightly (up to 2.6×), performance remains acceptable for real-world analytics, with execution still in seconds. Overall, the framework improves data sharing efficiency while reducing duplication and vendor lock-in.

(Sim, Carini and Tu) proposed an ontology-based federated data access framework to integrate human studies information across multiple distributed biomedical databases. The study addresses the problem of fragmented and siloed clinical research data, which limits large-scale analysis and evidence-based medicine. To overcome this, the authors developed an infrastructure that enables federated querying, allowing data to remain in its original sources while still being accessible for

unified analysis. The proposed system is built on three main components: the Ontology of Clinical Research (OCRe), which provides standardised semantic definitions for clinical concepts; an OCRe-derived data model that ensures consistency and supports data acquisition; and a query integrator that distributes queries across multiple databases and ontologies such as SNOMED and BioPortal. Overall, the approach demonstrates how federated data access can improve interoperability and enable large-scale biomedical research without requiring centralisation of sensitive data.

4.6. GDPR implications

(Karampela, Ouhbi and Isomursu) examined how the General Data Protection Regulation (GDPR) influences individuals' willingness to share personal and health-related data in healthcare settings. The study focuses on understanding user attitudes toward sharing different types of data, including health information, personal beliefs, consumption habits, and financial (wealth) data. Using a survey across four European countries with over 8,000 participants, the authors found that people are generally more willing to share health data and belief-related information compared to financial or wealth-related data, which are considered more sensitive. The study highlights that GDPR has significantly influenced data-sharing behaviour by increasing awareness of privacy rights and shaping users' trust in digital healthcare systems. Overall, the findings emphasise the importance of balancing data privacy regulations with the need for data sharing in healthcare innovation.
(Mourby, Mackey and Elliot) examined the implications of the GDPR for pseudonymised data in administrative data research in the UK. The study focuses on a key legal and technical issue: whether data that has been "pseudonymised" (e.g., key-coded or de-identified) should automatically be considered personal data under the GDPR, and how this affects data sharing in research contexts. The authors argue that GDPR does not necessarily expand the definition of personal data simply because data is pseudonymised. Instead, the critical test remains whether individuals can be identified using "means reasonably likely to be used," as outlined in Recital 26 of the GDPR. They explain that pseudonymised data can still be considered anonymous in certain contexts, especially when the risk of re-identification is very low or when accessed by third parties without identifying keys. However, they also highlight that within an organisation, pseudonymised data is still generally treated as personal data. Overall, the paper clarifies that pseudonymisation reduces risk but does not automatically remove data from GDPR scope.
(Biswal and Kulkarni) examined the implications of the GDPR on emerging technologies in healthcare, focusing on how data protection regulations affect the adoption and operation of modern digital systems in healthcare organisations. The study highlights that healthcare institutions increasingly rely on data collection and analytics to improve services, but this creates significant challenges in ensuring compliance with GDPR requirements. The authors explain that GDPR introduces strict rules for how personal data must be collected, processed, and stored, which directly impacts technologies such as artificial intelligence and data-driven healthcare systems. They emphasise that while GDPR compliance can be challenging for organisations, it is necessary to ensure data security, privacy protection, and long-term sustainability of digital healthcare services. The paper also suggests that organisations need better awareness and structured implementation strategies to align emerging technologies with GDPR principles.

4.7. Alignment with EU AI Act principles

The article by (Jonnala, Parida and Thomas) critically evaluates the EU Artificial Intelligence Act in relation to the emerging risks of generative AI systems. It finds that while the Act establishes a solid foundation for regulating AI through principles such as transparency, accountability, and risk-based classification, it does not fully capture or address the complexity and rapidly evolving vulnerabilities associated with generative AI. The study highlights gaps in practical implementation and enforcement guidance, particularly in areas like algorithmic fairness, explainability, and broader

socio-economic impacts. Overall, the authors conclude that although the EU AI Act is an important step toward responsible AI governance, it requires further refinement to remain effective in managing advanced AI systems.
The article by (Kwilinski and Reznik) examines the governance of artificial intelligence systems in the European Union and Ukraine, focusing on legal and institutional frameworks aligned with the EU AI Act. It explores how the EU's regulatory model promotes a balance between innovation and accountability through human-centred governance, digital ethics, and ESG principles. The study also highlights Ukraine's efforts to align its national AI strategies with EU standards as part of broader digital transformation initiatives. Overall, the authors conclude that effective AI governance requires coordinated legal, technological, and policy approaches to ensure responsible and sustainable AI development.
(Ceravolo, Damiani and D'Amico) proposes a structured Human Rights Impact Assessment (FRIA) framework (HH4AI) for evaluating AI systems under the EU AI Act. The framework is designed to identify and manage risks in AI systems by systematically assessing transparency, accountability, data governance, and human rights impacts across the AI lifecycle. It addresses the challenge of defining and regulating "high-risk" AI systems by providing a step-based methodology that supports compliance with EU AI Act requirements and strengthens ethical and legal oversight of AI deployment.
(Bernard) discuss how organisations are increasingly adopting AI governance and cybersecurity practices aligned with the EU AI Act, emphasising the need for strong oversight, transparency, accountability, and privacy protection in AI systems. The article highlights that frameworks like the EU AI Act and NIST AI Risk Management Framework are shaping how industries manage AI risks, particularly in security-focused applications, ensuring systems remain ethical, reliable, and compliant with emerging regulations.

5. Synthetic Data Generation for Multimodal Learning

5.1. Data Fusion and Synthetic Data Generation

Data is the fuel for AI, and in manufacturing quoting, relevant data can be particularly scarce and sensitive. The state-of-the-art is increasingly moving toward hybrid data strategies that combine real (domain-specific) data with synthetic data to train robust AI models. In cost estimation research, a notable example is the work of (Mandolini, Manuguerra and Sartini), who addressed the lack of historical cost records for custom-engineered products by generating a synthetic dataset. They built an automatic cost simulation tool to produce thousands of sample cost data points from 3D models, which then trained a machine learning model. This approach overcame sparse real data and resulted in a model that could predict costs within 7% error while saving significant time, and importantly, the model was explainable and updateable by design. In the realm of LLMs, synthetic data generation (e.g., via self-instructor prompt paraphrasing) has been used to fine-tune models for specific domains. For instance, to specialise a general language model for manufacturing, one might generate synthetic dialogues about quoting scenarios or translate existing RFQs into other languages to augment multilingual training. Companies like Paperless Parts hint at using proprietary data plus careful augmentation. Their Wingman AI was developed with a focus on secure training on customer data and presumably some synthetic augmentation for feature recognition (they emphasise it is powered by a proprietary secure AI model, using over 10,000 spec keywords, likely compiled from both real prints and simulated examples). Another example is the use of digital twins and simulation in manufacturing. By simulating AM builds and generating synthetic outcomes (e.g., print success/failure labels, cost under different machine settings), researchers can create datasets that inform AI models beyond what limited real experiments would allow.

The challenge with data fusion lies in ensuring quality and relevance. Domain-specific data (like past quotes, machine logs, customer RFQs) are highly valuable but often limited in quantity and variety, and subject to privacy constraints. On the other hand, synthetic data can be generated in large volume but may not capture all real-world intricacies. There is a risk models will learn artifacts of the synthetic process or that synthetic scenarios will not fully match what actual customers request. Current practices sometimes struggle with this balance. If not carefully managed, an AI might become over-fitted to synthetic patterns or, conversely, biased by a few real case outliers. Additionally, maintaining data security is paramount in B2B manufacturing. Real RFQs and CAD files contain proprietary information. Many state-of-the-art solutions avoid sending such data to public AI APIs. For example, Paperless Parts explicitly highlights that shops must be wary of how AI models are trained and deployed, implicitly promising that their in-house AI respects ITAR/export controls and other security needs. This indicates a limitation in using large pre-trained models: without fine-tuning or containment, they cannot incorporate proprietary data due to confidentiality. Finally, multilingual data fusion is an emerging concern. Assembling parallel corpora of RFQs in different languages might require synthetic translation, and ensuring the model performs equally well across languages needs careful curation.
Beyond SotA: This project introduces a hybrid data strategy that combines real domain-specific data with carefully generated synthetic datasets to address data scarcity, improve generalisation, and support multilingual quoting. We will develop a secure, simulation-driven data pipeline with parametric CAD variation, multilingual paraphrasing, and labelling to create diverse, high-fidelity training samples. Techniques such as contrastive learning and domain regularisation will ensure robust and privacy-preserving model performance across varied quoting scenarios.

5.2. Simulation-Based Data and Digital Twin

Simulation-based synthetic data is generated from explicit models of a system rather than from direct field observations. These models may be physics-based, rule-based, agent-based, or hybrid. Their common purpose is to reproduce system behaviour under controlled assumptions and generate data for training machine learning models. This approach is useful when real experiments are expensive, risky, slow, or operationally constrained. In such settings, virtual experimentation enables repeated scenario exploration without disturbing the physical system (Kantaros, Ganetsos and Pallis) (Liu and Istvan). In industrial contexts, this reduces data collection cost and supports the study of rare or safety-critical operating conditions.
A digital twin is a more tightly coupled and operational form of simulation. Unlike an offline simulator, it remains linked to a physical counterpart through ongoing data exchange. This link supports updating, monitoring, and decision-making over time. Recent digital twin literature treats this operational coupling as a defining feature of the concept. The virtual model should ingest data from the real system, remain sufficiently faithful for the intended task, and support actions such as prediction, diagnosis, optimisation, or control (Kantaros, Ganetsos and Pallis). The 2024 Nature Computational Science editorial also argues against treating a digital twin as an exact one-to-one copy of reality. Instead, it presents the twin as a fit-for-purpose computational asset. Its fidelity, complexity, and update frequency should match the target use case, available data, and computational constraints.
This distinction matters for synthetic data generation. A conventional simulator can produce synthetic observations. A digital twin extends that capability by combining simulation with current measurements, feedback loops, and targeted experimentation. Liu and David (2025) argue that this is especially relevant for modern AI systems, where limited data volume and quality often constrain training (Liu and Istvan). In their view, digital twins do more than generate data. They also help maintain data validity. If the simulator is outdated or cannot answer a query reliably, the link to the

physical system supports targeted re-observation and model refinement. This property makes digital twins useful in sim-to-real workflows because they support iterative calibration between simulated and observed behaviour.

From a data perspective, digital twins offer three main benefits. First, they enable controlled exploration of corner cases, including failure states and rare combinations of environmental variables. Second, they can generate synthetic sensor streams that remain anchored in process knowledge, physical constraints, or calibrated system behaviour rather than pure statistical imitation. Third, they support hybrid datasets that combine synthetic and real observations. For example, simulated data can support pretraining or stress testing, while real measurements support calibration and final adaptation (Kantaros, Ganetsos and Pallis; Liu and Istvan). This is relevant in domains such as manufacturing and autonomous systems, where collecting representative real data for all operating conditions is impractical.

Digital twins are not a substitute for real data in every setting. Their usefulness depends on validation, interoperability, uncertainty handling, and the cost of maintaining sufficient fidelity. Recent perspectives identify open problems in standards, scalable bidirectional data exchange, surrogate modelling, privacy, and human-in-the-loop operation. Kantaros et al. (2025) also emphasise scalability, uncertainty quantification, and heterogeneous data integration as unresolved barriers. The current literature therefore presents digital twins as evolving simulation infrastructures rather than perfect replicas. They are most effective when their scope, assumptions, and maintenance strategy are matched to the intended task.

5.3. Modality-Specific Synthesis - Images, Video, and 3D / Point Clouds

Synthetic data generation is strongly modality dependent. The representation, annotation cost, and realism criteria differ across images, videos, and 3D data. However, the main trade-off is similar across modalities. Synthetic pipelines offer controllability, scale, and automatic labelling, but they must still reduce the gap to real sensor data.

Synthetic images remain the most mature visual modality. They are comparatively easy to render at scale and can carry dense labels such as segmentation masks, depth, or pose with little additional annotation effort. In simulation-based pipelines, image synthesis benefits from explicit control over scene composition, lighting, camera placement, and object states. Martinez-Gonzalez et al. (2020) presents UnrealROX as a photorealistic virtual environment in which scene information is recorded per frame and replayed offline to generate raw data and ground-truth annotations for tasks such as semantic segmentation, object detection, and depth estimation (Martinez-Gonzalez, Oprea and Garcia-Garcia). This supports a broader move toward scene-first generation, where one reusable virtual environment serves multiple output types. The main advantage is geometric and semantic consistency across outputs. The main limitation is that transfer still depends on how well rendering, materials, and sensor effects match reality.

Synthetic video extends the same logic into the temporal domain. Compared with static images, video synthesis must preserve object identity, geometry, and appearance over time. It must also maintain plausible motion and temporal coherence. Virtual KITTI 2 is one example of this approach. Cabon et al. (2020) present it as a set of cloned driving sequences rendered in a virtual world with controlled variations in weather and camera configuration (Cabon, Murray and Humenberger). The dataset provides temporally consistent video data together with RGB, depth, class segmentation, instance segmentation, flow, and scene flow.

The current state of the art in synthetic point clouds is dominated by simulation pipelines that explicitly model geometry, sensor placement, and noise. Yaghi et al. (2025) describe recent LiDAR simulation work as a shift away from handcrafted virtual environments and toward more realistic and automated pipelines (Yaghi, Tekli and Kamradt). These pipelines use methods such as physics-based rendering, ray casting, and learned sensor-noise refinement. The main contribution of

3DGENie is to combine procedural layout generation, scene construction, asset randomisation, and synthetic sensor placement. This enables flexible 3D dataset generation for scenarios such as car assembly lines and autonomous driving. An important point is that realism is not only visual. It also depends on the statistical and structural properties of the sensor output. For this reason, 3DGENie emphasises configurable sensor simulation and alignment with real task conditions.

Hottong et al. (2024) make a related point from a different angle. They argue that the domain gap in synthetic point-cloud training is often driven not only by scene content but also by simulator parameters (Hottong, Sperling and Müller). They therefore propose automating the search for better domain settings during synthetic dataset creation. In this view, the generator itself becomes an optimisation target. This shifts synthetic data generation away from manual asset authoring and toward systematic calibration of the generation process. It also suggests that synthetic data quality should be judged by downstream task performance rather than by realism alone.

Alongside LiDAR-style simulation, 3D Gaussian Splatting has emerged as a complementary representation for synthetic data workflows. Chen and Wang (2026) describe 3DGS as an explicit scene model built from learnable 3D Gaussians rather than implicit neural radiance fields (Chen and Wang). This supports real-time rendering and improved editability. For synthetic data, the main value is that 3DGS can reconstruct detailed scenes from multi-view images or video and then support efficient re-rendering. This is useful for tasks such as novel-view synthesis, virtual reality, and autonomous driving. The same survey also identifies 3DGS as a bridge between graphics and simulation because Gaussian primitives can serve as rendering elements and carriers of physically meaningful structure. However, important challenges remain, including controllability, dynamic scene changes, physics-aware behaviour, and robustness under varying lighting conditions (Chen and Wang).

Taken together, the reviewed literature suggests a clear progression in modality-specific synthesis. Images are the most mature output. Videos are increasingly generated through reusable scene-centric pipelines. Simulation offers especially strong value for 3D and point-cloud synthesis because real acquisition and annotation are costly. Across modalities, the most promising direction is the use of shared scene representations that produce aligned images, video, depth, and 3D observations.

5.4. Generative AI Approaches for Time-Series Generation

Several representative machine learning approaches exist for time-series data generation, focusing on GAN-, VAE-, and diffusion-based models. These methods differ in how they model temporal dependencies and latent data distributions, ranging from adversarial learning to probabilistic latent-variable modelling and iterative denoising processes.

TimeGAN (Yoon, Jarrett and Van der Schaar) is a generative model for time series that combines adversarial training with recurrent components (RNN-based) to learn both feature distributions and temporal dependencies. It uses an embedding-recovery architecture to map sequences into a latent space and reconstruct them, with a supervised loss ensuring temporal consistency alongside adversarial training. It is evaluated on synthetic data (e.g., sine waves) and real benchmarks including stocks (Google stock, 2004- 2019), energy (UCI Appliances energy prediction dataset), and events (a large private lung cancer pathway dataset).

Conditional-TimeGAN1 (Liu, Ji and Chen) extends this model by adding conditional inputs to the generator and discriminator for condition guided time series synthesis. It enables class- or context-specific generation, particularly for non-intrusive load monitoring (NILM). The model is evaluated on synthetic single-appliance trajectory data and the REDD dataset, demonstrating improved

appliance-level time series generation and enhanced data augmentation for NILM tasks. Overall, Conditional-TimeGAN adapts TimeGAN to structured, domain-specific generation while preserving its ability to model realistic temporal dynamics.

TimeVAE2 (Desai, Freeman and Wang) is a variational autoencoder designed for multivariate time series generation, which encodes entire sequences into a structured latent space and reconstructs them through a neural decoder. It extends standard VAE formulations by incorporating time-series-specific inductive biases, enabling stable and interpretable generation through a probabilistic latent framework. The model is evaluated on four real-world multivariate time series datasets spanning domains such as finance, energy, and air quality. These include stock (price data from Yahoo Finance), energy (the UCI Appliances Energy dataset), and air (the UCI Air Quality dataset), along with synthetic time (sine waves).

Recent diffusion-based methods have become a strong approach for time series generation because they can model complex and diverse temporal patterns while training in a stable way. TimeLDM3 (Qian, Xie and Wan) proposes a latent diffusion framework for time series generation, where observed sequences are first encoded into a compressed latent space and then modelled via a denoising diffusion process, significantly improving computational efficiency while preserving high-fidelity temporal structure. It is evaluated on a mix of synthetic time series (sine waves) and real-world benchmarks, including financial time series (Google stock prices, 2004-2019), physical simulation data (MuJoCo multivariate dynamics), energy data (ETTh Electricity Transformer Temperature dataset), and health data (fMRI BOLD signals).

Diffusion-TS4 (Yuan and Qiao) proposes a diffusion-based framework for general time series generation with interpretability-aware constraints, where the diffusion process is designed to preserve meaningful temporal structure while enabling more controllable and structured synthesis across diverse domains. It is evaluated on a mix of synthetic and real-world benchmark datasets, including financial time series (Google stock prices, 2004–2019), electricity consumption data (ETTh Electricity Transformer Temperature dataset with load and oil temperature measurements), energy data (UCI appliance energy prediction dataset), physical simulation data (MuJoCo multivariate dynamics), health (fMRI BOLD signals), and synthetic sinusoidal signals (sine waves).

BRIDGE5 (Li, Huang and Xu) introduces a text-controlled framework for time series generation using diffusion models, where natural language descriptions guide the synthesis of temporal sequences. It employs a multi-agent LLM pipeline to bootstrap and refine text–time series pairs, addressing the lack of aligned training data for this task. Compared to latent diffusion methods such as TimeLDM, BRIDGE emphasises explicit natural language controllability and data construction via LLM-based augmentation.

5.5. Validation of Synthetic Data

The contemporary era of artificial intelligence is defined by foundation models pre-trained on internet-scale multimodal data, encompassing vision, audio, spatial networks, and text. As the observation space expands, the voracious data appetite of these architectures has exposed a critical limitation: the imminent depletion of organic, high-quality, and ethically sourced data. In response, the research community has aggressively pivoted toward Synthetic Data Generation (SDG).

By leveraging the generative capabilities of existing models to train subsequent generations, researchers can bypass traditional data walls and privacy constraints. Yet, the ingestion of synthetic data introduces a precarious dynamic often referred to as model collapse, where models recursively trained on synthetic outputs begin to experience performance degradation (Wen, Feng and Kang). Consequently, the operational challenge has fundamentally shifted. The industry is no longer constrained by the inability to generate data; rather, it is limited by the inability to systematically validate it. High-performance MSDG requires a delicate equilibrium: the synthesised data must exhibit enough *fidelity* to represent the target distribution accurately, while maintaining enough *diversity* to ensure the downstream model generalises beyond memorised modes (Pham, Nguyen and Le), (Pastrián, Hidalgo and Reyes).

As the volume of synthetic data scales, so does the sophistication of the artifacts it leaves behind. Standard transitional tampering artifacts have been superseded by complex structural distortions inherent to direct image synthesis. To counter this, late 2025 saw the rise of specialised multimodal models trained explicitly to not just detect, but linguistically explain these synthetic artifacts, bridging the gap between binary classification and human-interpretable forensic analysis (Wen, Feng and Kang).

As generation pipelines become unified, validation protocols must follow suit. The greatest vulnerability in legacy evaluation frameworks is the "Stitching Fallacy", the erroneous assumption that if synthetic text and synthetic tabular data are evaluated independently and both score high on fidelity, their combination is also high-fidelity. According to (Yuan, Atwal and Cheng), isolated unimodal evaluators consistently fail to detect critical cross-modal logical contradictions. For example, a tabular health record synthesising a diagnosis of "Diabetes" paired with an unstructured text summary discussing a "Fractured Femur" will pass standard Natural Language Processing NLP metrics (like BERTScore) and tabular statistical tests but fundamentally fail clinical validity SynEval resolves this by mapping discrete tabular features and continuous text embeddings into a singular, joint geometric space, allowing for the explicit modelling of the joint distribution of multimodal data. (Yuan, Atwal and Cheng). In the purely visual and spatial domains, benchmarking has moved away from binary "real vs. fake" classification classifiers, which are notorious for demographic and topological biases. The current standard requires diagnostic explainability. Frameworks like FakeClue and the accompanying FakeVLM architecture evaluate synthetic data by forcing detection models to generate natural language explanations for the artifacts they identify. To counter the inherent biases or "hallucinations" of a single detection model, FakeClue aggregates annotations from multiple high-performance LMMs, establishing a rigorous benchmark that categorises (Wen, Feng and Kang)

A major operational bottleneck in multimodal learning is synthesising edge-case data that adheres to strict physical or domain-specific laws. When generating interleaved sensor data, physiological metrics, and unstructured text, unconstrained models frequently hallucinate physically impossible scenarios. To resolve this, researchers have introduced multi-agent scaffolding frameworks. In these advanced architectures, generative agents do not output directly to a dataset. Instead, they submit their multimodal drafts to a committee of evaluator agents equipped with deterministic, rule-based soundness checkers. According to a recent 2026 pre-print on synthesising high-frequency kinematic sensor data and physiological markers (Al-Kabbany and Kassem), implementing a multi-agent LLM architecture that evaluates synthetic samples against rigid domain rules (e.g., physiological limits) ensures that the output dataset maintains high fidelity without requiring costly human-in-the-loop annotation. This process generates validated "Question-Context-Answer" triplets that strictly map raw sensor spikes to evidence-based outcomes, effectively bridging raw data engineering with practical, constraint-heavy sciences (Al-Kabbany and Kassem).

If naive synthetic data is injected into an active learning pipeline, models are highly susceptible to "bias inheritance." Because foundational generative architectures are heavily biased toward the modes of their organic training distributions, synthetic augmentation can inadvertently suppress

demographic and topological minorities, severely damaging algorithmic fairness. To counteract this, the research frontier has developed multimodal fairness frameworks that explicitly decouple the fidelity-diversity trade-off. Methodologies like Fair Diffusion (Friedrich, Brack and Struppek) force the generative process to sample from the tail ends of the latent distribution. By actively mapping and over-sampling underrepresented demographic or structural features during synthesis, these frameworks guarantee that the resulting downstream models achieve demographic parity and equalised odds, often outperforming models trained exclusively on organically sourced but naturally biased datasets (Friedrich, Brack and Struppek).

The modern training loop for Large Multimodal Models now fundamentally relies on adversarial self-correction. Artefact detection is no longer just an isolated benchmark evaluated at the end of the pipeline; it is a continuously updated penalty mechanism during the synthesis phase. Models like FakeVLM not only distinguish real from synthetic data but dynamically generate fine-grained, natural language artifact explanations (Wen, Feng and Kang).

By looping these diagnostic explanations back into the generative model, the pipeline creates a closed-loop refinement environment. However, rather than waiting to adjust model weights in subsequent training epochs, this correction happens dynamically at inference time. As demonstrated by Chern et al. (Chern, Hu and Chern), Large Multimodal Models can construct intermediate visual drafts and explicitly critique their own structural or logical flaws through text. By feeding these natural language critiques directly back into the immediate conditional prompt, the generator iteratively refines its output. This "generate-critique-refine" loop ensures that specific semantic and structural signatures are corrected on the fly, guaranteeing that only high-fidelity, artifact-free samples are finalised for the downstream synthetic dataset (Chern, Hu and Chern).

As we project into the next three to five years of Multimodal Synthetic Data Generation (MSDG), the field is preparing for an inflection point driven by the "recursive generate-train loop." Foundational models are increasingly being fine-tuned on the synthetic exhaust of their peers, fundamentally altering the statistical landscape of the Internet's data ecology.

The most critical area of near-future research is understanding and mitigating multimodal model collapse. Historically, model collapse in unimodal generation was characterised by an inevitable loss of variance. The generator would regress to the mean, producing highly homogenised outputs. However, recent empirical breakthroughs demonstrate that multimodal collapse operates under entirely different statistical mechanics. When observing the generate-train loop between Vision-Language Models (VLMs) and text-to-image diffusion models, researchers have found that variance does not necessarily decrease. Instead, as demonstrated by (Hu, Rostami and Thomason), the recursive training of synthetic data in VLMs often leads to *increased* variance in specific tasks like image captioning, alongside an anomalous initial improvement in vision-language alignment before ultimate structural degradation. The future roadmap heavily relies on mapping these specific multimodal degradation patterns, ensuring that increased decoding budgets and variance controllers are implemented to build datasets robust enough to withstand recursive epochs.

To sustain the fidelity-diversity equilibrium over infinite scaling laws, the next generation of MSDG pipelines will likely abandon purely statistical validation in favour of neuro-symbolic and cryptographic interventions:

- Neuro-Symbolic Constraint Systems: Future multimodal architectures will embed symbolic logic engines directly into the diffusion latent space. This will explicitly prohibit the generation of physical or logical impossibilities (e.g., shadows pointing toward a light source in an image, paired with audio indicating a different environment), resolving the cross-modal stitching fallacies that purely statistical models fail to catch (Yuan, Atwal and Cheng).
- Immutable Latent Watermarking: As LMM-based detection benchmarks like FakeVLM map the artifacts of synthetic data (Wen, Feng and Kang), (Rebuffi, Tran and Lacatusu)[OBJ].

The transition from data scarcity to synthetic abundance has definitively shifted the engineering bottleneck from generation to validation. Navigating the fidelity-diversity trilemma requires treating synthetic data not as a cheap substitute for organic data, but as a deeply complex, multi-objective optimisation problem. By enforcing unified geometric distribution frameworks, utilising multi-agent artefact detection, and embedding strict fairness constraints, the research community is successfully forging a path where LMMs can safely and iteratively self-improve. The future of multimodal AI depends entirely on the rigorous auditing of its synthetic diet.

6. State of the Art in Multimodal Data Fusion

6.1. Multimodal Processing of Text and 3D Files

Interpreting an RFQ for manufacturing often requires understanding information from multiple modalities, free-form text (emails, PDFs) and 3D part data (CAD files, technical drawings). Thus, combining natural language processing (NLP) and 3D shape analysis is crucial. State-of-the-art systems are beginning to tackle such multimodal fusion. For instance, the DigiFabster AI agent described above links an email's text to its CAD attachment. It reads the email for quantities or deadlines and simultaneously analyses the CAD model's geometry before producing a quote. Paperless Parts' Wingman similarly processes text in drawings or BOMs together with the 3D model to highlight important features, e.g. identifying a material call-out on a PDF drawing and matching it to the CAD file's metadata. Under the hood, these systems use a combination of NLP and CAD parsing algorithms.

In research, PointNet (Qi, Su and Mo), PointNet++ (Qi, Yi and Su) and MeshCNN (Hanocka, Hertz and Fish) have enabled deep learning on 3D shapes. Recent work like M3D-Bench (M. Li, X. Chen and C. Zhang) attempts to extend large language models to handle 3D data by creating instruction-following datasets that intermix text, images, and 3D objects. Such efforts recognise that current multimodal LLMs excel at 2D image understanding but lack native capability for 3D geometry, spurring development of 3D-aware AI assistants. Another example is a hybrid system by (Zhang and Zhao) that presents a web-based manufacturability advisor taking both textual inputs and 3D models. It applies a machine learning model to evaluate a part's printability from its CAD file and also uses rule-based logic to interpret design requirements, effectively functioning as an expert system reading design documents and geometries together.

Key limitations persist in today's multimodal processing. Seamless fusion of text and 3D information is hard. Most solutions process each modality separately and then merge results heuristically. General-purpose large multimodal models have not yet been trained to take a raw CAD model as input alongside a prompt. Thus, companies resort to building custom pipelines. This can lead to brittleness when RFQ instructions contradict CAD details or when dealing with varied file formats. Additionally, data representation is an issue as 3D CAD data is complex (meshes, B-reps, point clouds) and not as readily interpretable by transformers as text or images. Specialised 3D neural networks exist for shape classification or feature detection but integrating them with NLP models in real-time quoting scenarios is non-trivial. Current industry tools that do combine modalities (like Wingman) focus on specific, known extraction tasks rather than truly understanding the part in a holistic sense. As a result, there can be gaps, for example, a system might extract part dimensions from a CAD file and material from text but miss implicit interactions (such as a tolerance note in the email that affects manufacturability of a small feature in the CAD). Finally, multilingual support in multimodal contexts is rare. A model may parse geometry regardless of language, but interpreting RFQs in, say, German or French requires multilingual NLP, which many custom tools do not possess.

Beyond SotA: We aim to establish a novel multimodal AI framework that performs deep fusion of textual, 3D and 2D data streams through early and intermediate architectures. Going beyond the current practice of loosely linking NLP and CAD analysis, our system will experiment with shared representation spaces where language-derived quoting parameters (e.g., “tight tolerance on slot”) and geometry-derived descriptors (e.g., overhang complexity) are embedded and reasoned over jointly. This includes investigating transformer-based multimodal models that natively ingest mixed-modality inputs (e.g., text prompts and point clouds) and context-aware alignment techniques to resolve contradictions or missing links between modalities.

7. Public and Benchmark Multimodal Datasets

The effectiveness, reliability, and emerging capabilities of a Multimodal large language model (MLLM) are inextricably linked to the quality, scale, architectural formatting, and structural diversity of its training and evaluation datasets. As the industry moves toward omni-modal perception, the literature emphasises that achieving deep cross-modal alignment requires a transition from simplistic image-text pairs to massive, heterogeneous data streams.
In this section, as shown in Table 1, we examine public datasets. In selecting the dataset, a balanced coverage strategy is followed to include different modalities, different sizes and scales, different objectives, and different learning paradigms.

Table 1. Comparison of Public Multimodal Datasets

Dataset	Modalities	Size	Domain	Key Features
LAION-400M (Schuhmann, Vencu and Beaumont) & LAION-5B (Schuhmann, Beaumont and Vencu)	Text, Image	400 million pairs & 5.85 billion pairs	General-purpose dataset	CLIP-filtered image-text pairs
InternVid (Wang, He and Li)	Text, Video	7 million videos & 234 million clips & 4.1 billion words	General-purpose dataset	Segment-level temporal descriptions.
ShareGPT4V (Chen, Li and Dong,)	Text, Image	102,000 pairs	General-purpose dataset	High-density, multi-turn conversation
EuroBlocks Instruction Datasets (Martins, Alves and Fernandes) (Ramos, Alves and Gisserot-Boukhlef)	Instruction & Response pairs (Multilingual)	10.6 millions instruction–response pairs & 24 EU languages	General-purpose dataset	Legally safe, synthetic data for EU AI Act compliance.
nuScenes (Caesar, Bankiti and Lang)	LiDAR, Radar, Camera	1.4 million 3D boxes	Autonomous Driving	360° sensor coverage with

Dataset	Modalities	Size	Domain	Key Features
				manual 3D annotations.
MultiIoT (Mo, Morency and Salakhutdinov)		12 Data Modalities such as video, audio, images, and LiDAR	1.15 million samples	IoT / Smart Devices	Time-series-based multimodal IoT dataset
RadioML (O'Shea, Roy and Clancy)		In-phase and Quadrature (IQ) Signals & 24 modulation types	2.5 million signals	Telecommunications	1D signals converted to 2D waterfall spectrograms
Geolife (Zheng, Xie and Ma)		GPS Trajectories	17,621 GPS trajectories	Mobility Analysis	A large-scale GPS-based spatiotemporal dataset
SciTabAlign (Ho, Wu and Kumar), (Ho, Kumar and Wu)		Text, Tables, Charts	136 tables and 372 claims	Scientific Reasoning	Maps logically equivalent but visually distinct representations.
ChartMimic (Yang, Shi and Liu)		Chart, Instruction, Code	4800 Triplets	Scientific Reasoning	Enable models to understand, interpret, and produce data visualisations.

LAION-400M and **LAION-5B** are open datasets containing image-text pairs collected from the internet. LAION-400M was released in 2021 and contains 400 million English image-text pairs. The image-text pairs were extracted from the Common Crawl web dataset and were collected from randomly selected web pages crawled between 2014 and 2021. All data was filtered using OpenAI's CLIP tool by calculating the cosine similarity between text and visual embeddings and removing pairs with a similarity score below 0.3. LAION-5B, released in 2022, contains 5.85 billion multilingual image-text pairs and was filtered using the CLIP tool. It is possible to use datasets by creating subsets. Since datasets contain raw internet data, they may contain bias and harmful content; therefore, a rigorous cleaning process is essential for industrial use.

InternVid, released in 2023, is a dataset consisting of over 7 million video-text (caption) pairs, collected from YouTube, covering 16 topics, with an average duration of 6.4 minutes, and totalling approximately 760,000 hours of content. The data collection process is based on two main approaches. First, popular channels across different categories, such as news and gaming, were analysed, and trending or highly rated videos from these channels were selected, resulting in approximately 2 million videos. Second, a list of verbs describing various actions and activities was compiled; videos ranking at the top of search results using this list were selected, adding an

additional 5.1 million videos to the dataset. Captions were generated automatically, so the noise ratio may be high.

ShareGPT4V is a dataset based on synthetic data containing approximately 100,000 images and over 1.2 million multi-turn dialogues related to these images. A multi-turn conversation refers to a dialogue structure that progresses over multiple steps, with the model maintaining context by taking previous interactions into account. It is designed to enable multimodal models to engage in human-like, context-aware, and instruction-focused communication through images, prioritising explanatory richness over sheer volume.

The EuroBlocks Instruction datasets are large-scale, multilingual, supervised fine-tuning datasets that provide 10.6 million instruction-response pairs across general, coding, mathematics, and STEM domains. The datasets consist of 60\% English general tasks, 20\% multilingual general tasks, and 20\% STEM/coding/math-related tasks. In particular, they have been used in the development of EuroLLM-22B and other related large language models.

nuScenes is a dataset developed for autonomous driving research, comprising 1.4 million accurately labelled 3D, multi-sensor, multimodal, and real-world data points derived from 40,157 key frames captured using 6 cameras, 1 LiDAR, and 5 RADAR sensors. The nuScenes ecosystem also includes the nuImages dataset, comprising 93,000 two-dimensional images captured by the 6 cameras, and the nuReality dataset, which contains both simulation and real-world data.

MultiIoT a time-series-based multimodal dataset that aggregates multi-sensor data from IoT devices. The dataset includes 1.15 million samples across 12 different modalities—such as video, audio, images, and LiDAR—and covers 8 distinct scenarios, including the detection of human posture, gaze, activities, and gestures, as well as the detection of physical objects' touch, contact, posture, and 3D structure.

RadioML is a dataset created using a Software-Defined Radio (SDR) system—commonly employed in research on Automatic Modulation Classification (AMC) in wireless communications—that considers numerous radio channel characteristics closely resembling real-world radio signals. The dataset includes 24 types of analogue and digital modulation and contains 2.5 million signals.

Geolife is a dataset collected to analyse human mobility and daily life behaviours, representing a total distance of over 1.2 million kilometres, a total duration of 48,203 hours and comprising 17,621 GPS trajectories.

SciTabAlign is a multimodal (text + structured data) dataset that demonstrates the semantic relationships (alignments) between statements in scientific texts and tables, utilising both table and text data together. Since SciTabAlign presents tables in a text-based format, it is suitable for generating related graphics and contains 136 tables and 372 claims. It is an expanded version of the SciTab dataset.

ChartMimic is a multimodal dataset comprising 4,800 human-curated triplets (chart, instruction, code) that pair actual graphical images found in scientific articles across various fields with their textual descriptions. These graphics encompass 18 standard types and 4 advanced types and are divided into 201 subcategories.

We can summarise the findings from our analysis of public multimodal datasets as follows.

- Across modalities, achieving full representation is very difficult due to losses in mappings. Unlike in the real world, a complete homomorphism has not been observed between visuals, audios, and texts.
- Labels linking two modalities are assigned in scenarios involving human subjects. Consequently, the labels in the dataset are the result of a subjective decision. Attention should be paid to this issue, especially in multi-dialogue datasets.
- As the number of modalities increases, the sample density within each modality decreases. This suggests a potential trade-off between breadth and depth.

- Some datasets have multiple versions. Multiple modalities within a dataset may not remain up-to-date simultaneously, and an asymmetric distribution shift may occur. The results of a model trained on these datasets may vary depending on the modality.
- While these datasets are publicly available, their representativeness is debatable. The recorded data consists of public information. It should also be noted that there are cases where data cannot be recorded.

These findings will undoubtedly shape the strategy for use cases in CLEAR projects. Subjective labelling, modality weighting, the balance between scope and depth, versions, and the timeliness of the data, as well as their representativeness, are factors that must be considered.

Table 2. This table maps the public datasets to CLEAR use case domains (software engineering, 3D, time series and IoT). Further, the table also analyse the relevance of the dataset to the CLEAR project.

Dataset	CLEAR Domain	Relevance to CLEAR project
LAION-400M (Schuhmann, Vencu and Beaumont) & LAION-5B (Schuhmann, Beaumont and Vencu)	No direct relevance	The dataset contains images and captions from scrapped from the internet. There is no direct relevance to the CLEAR project. However, it can still provide minor, indirect value as a foundational pre-training source for general visual-textual intelligence.
InternVid (Wang, He and Li)	No direct relevance	The dataset is video text pairs with videos scrapped from YouTube. While it has no direct relevance to CLEAR use cases, the dataset can still guide developments in CLEAR's agriculture use case from the Estonian sub-consortium.
ShareGPT4V (Chen, Li and Dong,)	No direct relevance	Dataset of synthetic image-description pairs with no direct relevance to the CLEAR project. However, it can still provide minor, indirect value as a foundational pre-training source for general visual-textual intelligence
EuroBlocks Instruction Datasets (Martins, Alves and Fernandes) (Ramos, Alves and Gisserot-Boukhlef)	No direct relevance	Instruction response pairs that can guide instruction tuning of LLMs on multilingual data. However, no direct relevance of the dataset to the CLEAR project and the dataset lacks multimodality.
nuScenes (Caesar, Bankiti and Lang)	3D and Time Series	The dataset contains multimodal data from LiDAR, Radar and Camera. Could be highly relevant in benchmarking, finetuning and testing approaches in CLEAR domains i.e., 3D data and time series analysis.
MultiIoT (Mo, Morency and Salakhutdinov)	Time Series and IoT	A multi sensor IoT data with time series characteristics. Highly relevant for timeseries analysis and IoT domains within the CLEAR project for benchmarking and training.

Dataset	CLEAR Domain	Relevance to CLEAR project
RadioML (O'Shea, Roy and Clancy)	Indirectly relevant for IoT communication	A multi modulation radio frequency dataset for transmission identification tasks. The dataset can be partially relevant to the CLEAR project where IoT devices communicate over radio.
GeoLife (Zheng, Xie and Ma)	No direct relevance	GPS trajectories collected over 182 unique users with timestamps and transport mode labels. While there is no direct relevance to CLEAR domains, the dataset can guide data collection and analysis in CLEAR tasks with spatiotemporal nature, such as, emergency calls and agriculture use case.
SciTabAlign (Ho, Wu and Kumar), (Ho, Kumar and Wu)	Partially relevant for software engineering	The dataset contains claims against tabular data that can be used for claim conformance. Such dataset can be relevant in benchmarking and can guide approaches with software verification for conforming/verifying requirements against tabular signal data.
ChartMimic (Yang, Shi and Liu)	Partially relevant for software engineering	The figure, instruction, code triplets of the dataset could enable benchmarking and multimodal reasoning over visualisation-heavy software systems.

There exist well-established domain benchmarks for time-series forecasting (Table 3), including the **GluonTS benchmark** collection (Alexandrov, Benidis and Bohlke-Schneider) and the **Monash Time Series** Forecasting Archive (Godahewa, Bergmeir and Webb). These benchmarks span multiple real-world domains such as energy, transportation, environment, and weather and are highly relevant for the CLEAR project. The Electricity dataset contains hourly electricity consumption records of 321 clients from 2012-2014. The Solar dataset consists of a 137 hourly time series of solar power production in Alabama during 2006. The Wind dataset provides high-frequency wind power generation data recorded every 4 seconds starting from August 2019. The Traffic dataset includes daily freeway lane occupancy rates in the San Francisco Bay Area over 15 months. The Air Quality dataset contains hourly pollutant measurements from 59 monitoring stations in Beijing and London between 2017 and 2018. The Temperature dataset comprises daily mean temperature observations from 422 weather stations across Australia between 2015 and 2017. These datasets are widely used due to their diversity in temporal resolution, sampling frequency, and application domains.

Table 3. Benchmarks for time-series forecasting

Time-Series Dataset	Modalities	Size	Domain	Key Features
GluonTS and Monash (Alexandrov, Benidis and Bohlke-Schneider) (Godahewa, Bergmeir and Webb),	Time	Medium-scale (varies per dataset)	Electricity, Solar, Wind, Traffic, Air Quality, Temperature	Standard forecasting benchmarks
Time-MMD (Liu, Xu and Zhao)	Time + Text (real)	Large-scale (24+ years, multi-domain)	Agriculture, climate, economy, energy, environment,	Multimodal time-series with aligned textual context

Time-Series Dataset	Modalities	Size	Domain	Key Features
			health, security, social good, traffic	
TSQA (Kong, Yang and Hwang)	Time + Text (real + synthetic)	~200K QA pairs	Healthcare, finance, energy, traffic, IoT, environment, human activity, web, AIOps	Multi-task QA benchmark
CityEval (Feng, Liu and Du)	Time + Text + Image	Large-scale (multi-city, spatio-temporal)	Urban systems (mobility, infrastructure, activity)	Spatio-temporal reasoning; combines spatial, temporal, semantic context

The **Time-MMD dataset** consists of a large-scale collection of multimodal time-series data spanning 9 domains, including agriculture, climate, economy, energy, environment, health, security, social good, and traffic. It is built from real-world datasets covering more than 24 years of observation. The dataset integrates numerical temporal signals with aligned textual data from search and report sources, enabling the use of contextual information beyond raw time-series values. It was introduced to support tasks like forecasting and benchmark multimodal time-series learning, showing improved performance over unimodal approaches.

The **TSQA dataset** consists of around 200k question–answer pairs covering 12 domains, including healthcare, finance, energy, traffic, environment, IoT, nature, transport, human activities, machine sensors, AIOps, and the web. It is organised into five task categories: forecasting, imputation, anomaly detection, classification, and open-ended reasoning. Within the open-ended reasoning subset, it further contains 6,919 true/false questions, 11,281 multiple-choice questions, and 12,510 open-ended questions, thereby providing a wide variety of question formats for comprehensive evaluation across diverse real-world settings.

CityEval is a spatiotemporal urban dataset that incorporates time-evolving city-level signals across multiple regions. It captures temporal dynamics of urban systems such as mobility, activity, and infrastructure-related indicators, while also preserving spatial structure across cities or locations. The dataset is designed for evaluating models on urban understanding tasks that require reasoning over both temporal evolution and spatial context, rather than representing a pure time-series benchmark.

7.1. Multimodal Datasets for Software Testing and Systems Verification

Several of CLEAR's use cases concern software-intensive systems, for which the relevant multimodal datasets differ from the general-purpose corpora discussed above. This subsection complements the general datasets of Table 1 and their CLEAR-domain mapping in Table 2 by reviewing the state of the art and state of practice in multimodal datasets specific to software testing and systems verification, organised by the kind of testing or verification work each family of datasets enables.

State of the art versus state of practice.

Multimodal models are increasingly evaluated on their ability to reason over diagrams, tables, charts, scientific figures, documents, screens, and maps, rather than simply recognize objects. Benchmarks such as MMMU (Yue, Ni and Zhang), MathVista (Lu, Xu and Liu), MMBench (Liu, Duan and Zhang), and SEED-Bench (Li and Wang) reflect this shift toward reasoning-focused evaluation. The state of practice is more conservative, and for good reason. Most teams do not build LAION- or CLIP-scale corpora from scratch. Instead, they combine existing public datasets, private domain data, screenshots, logs, documents and sensor streams, foundation-model-generated labels, and a small but crucial loop of human validation. The pattern repeats across domains. In autonomous driving, multimodal practice is mature and highly engineered: synchronised, calibrated camera, LiDAR, radar, GPS/IMU and map data underpin datasets such as nuScenes (Caesar, Bankiti and Lang), the Waymo Open Dataset (Sun, Kretzschmar and Dotiwalla), Argoverse (Chang, Lambert and Sangkloy), and BDD100K (Yu, Chen and Wang). In healthcare, privacy and clinical validation dominate; MIMIC-CXR (Johnson, Pollard and Berkowitz), ROCOv2 , VQA-RAD (Lau, Gayen and Ben Abacha) and PathVQA (He, Zhang and Mou) pair medical images with reports, captions, concepts or clinician-style questions, but under strict access controls. In software testing and UI automation, multimodal datasets are emerging but not yet mature. The practical modalities are screenshots, DOM or view-hierarchy trees, OCR text, requirements, test steps, logs, and sometimes source code.

GUI and screen-understanding datasets.

Modern testing increasingly needs to understand what is on a screen — widgets, layouts, states and flows. These datasets teach exactly that (Table 4).

Table 4. Multimodal datasets for GUI and screen understanding

Dataset	Modalities & scale	Relevance to testing & verification
Rico (Deka, Huang and Franzen)	Android screenshots, view hierarchy, text, interaction traces; 66k+ screens, 9.3k apps, 27 categories	Foundational screen-understanding corpus, supporting GUI test generation, visual regression, locator and oracle work, and accessibility checks. Its main limitation is older app design.
MobileViews (Gao, Zhang and Chen)	Screenshot – view-hierarchy pairs; 600k+ pairs from 20k+ modern apps (later extended past 1M)	A modern successor to Rico for AI-assisted mobile testing: screen-state understanding, form and field detection, and visual oracles. Best paired with action traces for end-to-end bugs.
Screen2Words (Wang, Li and Zhang)	Screenshot, view hierarchy, app metadata, NL summaries; 112k summaries over ~22k Rico screens	Generates readable screen descriptions for test reports and test-step documentation; it describes screens but does not verify behaviour.
Widget Captioning (Li, He and Zhou, Widget Captioning: Generating Natural Language Description for Mobile User Interface Elements)	Screenshots, cropped widgets, view hierarchy, NL captions (built on Rico)	Connects visual widgets to functional meaning, enabling semantic locators that replace brittle XPath / resource-ids and flag missing labels.
ScreenQA (Hsiao, Li and Liu)	Screenshot + question + answer + UI bounding boxes; ~86k QA pairs over Rico	Enables visual assertions (“Does this screen show payment succeeded?”) instead of hard-coded text checks. It is a QA benchmark, not executable tests.

ScreenAI (Baechler, Ghosh and Haurilet)	Screenshots, UI annotations, QA pairs, generated screen descriptions (Screen Annotation, ScreenQA-Short, Complex ScreenQA)	Close to real testing questions: element presence, layout correctness, chart/table rendering, and UI navigation. Highly reusable for AI-assisted GUI testing.

GUI grounding and element localization.

Automated tests must know where to click, type, scroll or verify. Grounding datasets in Table 5 connect an instruction to a precise location on screen, which is the antidote to brittle locators.

Table 5. Datasets for GUI grounding and element localization

Dataset	Modalities & scale	Relevance to testing & verification
ScreenSpot (SeeClick) (Cheng, Sun and Chu)	Screenshot + NL instruction + target element box; 1,200+ instructions across iOS, Android, macOS, Windows, Web	Replaces fragile locators with visual-semantic grounding (“click Continue”), directly targeting locator maintenance — a major cost in automation.
ScreenSpot-Pro (K. Li, Y. Chen and Y. Wang)	High-resolution desktop screenshots + grounding instructions; 1,581 instructions, 23 professional apps, 3 OSes	Reflects professional, high-resolution enterprise tools, making it more relevant to desktop verification than mobile-only sets.
UGround (Gou, Chen and Fan)	Screenshots + referring expressions + element coordinates; ~10M elements over 1.3M screenshots	Among the largest grounding corpora; supports cross-platform visual click automation and test migration. It tells where to act, not whether output is correct.

Mobile GUI action and task completion.

Because executing a test is a sequence of UI actions, datasets that pair instructions with screen states and actions transfer naturally to automated GUI testing (Table 6).

Table 6. Datasets for mobile GUI action and task completion

Dataset	Modalities & scale	Relevance to testing & verification
PixelHelp / Seq2Act (Li, He and Zhou, Mapping Natural Language Instructions to Mobile UI Action Sequences)	NL instruction + mobile screen + UI objects + action sequences	Turns manual test steps (“open settings, enable dark mode”) into executable UI actions; instruction-following rather than bug detection.
MoTIF (Burns, Tan and Ghosh)	NL command + app environment + screen states + actions + feasibility labels + follow-up questions	Handles ambiguous or infeasible test steps (“delete the saved card” when none exists) — valuable for exploratory and negative testing.
Android in the Wild (AITW) (Rawles, Li and Wen)	NL instruction + Android screenshots + human action demos; 715k episodes, 30k instructions, 8 device types	Trains agents to execute tasks from natural language and tests robustness across devices and versions. It lacks view hierarchies, limiting precise UI reasoning.
AndroidControl (Li, Zhang and Wang)	Screenshots + demos + high- and low-level instructions +	Bridges high-level scenarios and low-level actions — useful because testers write

	actions; 15,283 demos, 14,548 tasks, 833 apps	scenarios while automation needs concrete steps.
AMEX (Chai, Zhang and Li)	High-res screenshots + element grounding + screen/element descriptions + multi-step instructions; 104k+ screenshots, 110 apps, ~13 steps each	Multi-layered (grounding, description, multi-step execution), covering much of what an AI tester must do on mobile. Mainly mobile-focused.
GUI Odyssey (+ CoM) (Lu, Chen and Zhang)	Screenshots + actions + goals + history across hundreds of apps; the CoM variant adds 111k chain-of-memory pairs	Targets cross-app, system-level navigation ("receive an OTP, enter it in a banking app, save the receipt to email") and memory-aware workflows.
OmniGUI (Henry, Lin and Zhu)	Static images + synchronous audio + video clips + action steps; 709 expert episodes, 2,579 steps, 29 apps	Points to dynamic multimodal UX testing where correctness depends on video, audio and screen together — frozen animations, missing sounds. New and small.

Web, browser and desktop automation.

These benchmarks evaluate agents performing real tasks in real environments — the closest analogue to end-to-end and system testing, especially where success is verified by outcome rather than by a single click (Table 7).

Table 7. Benchmarks for web, browser and desktop automation

Dataset	Modalities & scale	Relevance to testing & verification
Mind2Web family (Deng, Gu and Zheng) (Xue, Zheng and Wang)	NL task + HTML/DOM + actions + screenshots (Multimodal-Mind2Web); 2,350 tasks, 137 sites; Online-Mind2Web evaluates live sites	Close to automated web testing — multi-step navigation, form filling, verification. Online evaluation reflects changing real websites, not frozen snapshots.
WebArena (Zhou, Xu and Zhu)	NL task + self-hosted web environment + browser state + actions + backend verification	Very close to system testing: success is checked by whether the task truly completed (e.g., create, assign and comment on an issue), not just clicks.
VisualWebArena (Koh, Lo and Jang)	Web screenshots + text + NL instructions + actions; 910 visually grounded tasks	Best when a web test needs visual cues absent from the DOM — selecting by image, reading layout, or comparing visible items.
VideoWebArena (Jang, Zhang and Yang)	Video + text + images + web actions	Tests interfaces where tutorials, animations or video content drive the workflow and require memory and retrieval.
WorkArena / BrowserGym (Drouin, Gasse and Caccia)	Browser UI + NL task + multimodal observations + actions; 33 ServiceNow tasks	Enterprise workflow verification: knowledge-based search, form filling, catalogue ordering, and ITSM ticket operations.
OSWorld (Xie, Zhang and Chen)	Desktop screenshots + OS/app state + files + actions + execution-based evaluation; 369 real computer tasks	Among the most relevant for system-level automation spanning multiple desktop apps and files (e.g., update a spreadsheet, export a PDF, upload it).
GUI-World (D. Chen, Y. Li and Z. Wang)	GUI videos + keyframes + captions + QA + screenshots across 6 settings (desktop, mobile, web, software, XR)	Tests dynamic GUIs where a static screenshot is insufficient — loading states, transitions, animations, and multi-step history.

GUI-360 (Mu, Li and Chen)	Full-res screenshots + accessibility metadata + goals + trajectories + reasoning traces (successful and failed); 1.2M+ action steps	Office and enterprise verification: parse a screen, ground an instruction, predict the next action, follow a workflow, and recover from failures.

Webpage, front-end and UI-code datasets.

Front-end testing compares visual design, rendered UI and underlying code. These datasets connect screenshots to structure and to generated code, which maps onto visual regression and fidelity checks (Table 8).

Table 8. Datasets for webpage, front-end and UI-code fidelity

Dataset	Modalities & scale	Relevance to testing & verification
WebSRC (Chen, Xie and Li)	HTML source + screenshots + metadata + QA; ~400k QA pairs over 6.4k pages	Validates structural and visual page understanding ("is the shipping price under the right product?"), combining DOM and screenshot evidence.
Design2Code (Si, Yu and Chen)	Webpage screenshot + generated code + visual-comparison metrics + human evaluation; 484 real pages	Screenshot-to-code fidelity maps directly to front-end regression testing and screenshot-based oracles.
Web2Code (Yun, Lee and Kim)	Webpage images + instructions + HTML code + QA pairs	Tests whether models can understand and reconstruct interfaces, connecting to UI testing and visual regression.

Bug reports, test reports and software maintenance.

This family is closest to classical testing: it contains bug reports, test reports, screenshots, code and user feedback, and includes the rare datasets that encode an actual testing oracle (Table 9).

Table 9. Datasets for bug reports, test reports and software maintenance

Dataset	Modalities & scale	Relevance to testing & verification
ReCoDe (Yu, Wang and Zhang, Mobile App Crowdsourced Test Report Consistency Detection via Deep Image-and-Text Fusion Understanding)	Crowdsourced report text + app screenshots; 22k+ reports (only ~18% found consistent)	Checks whether a screenshot actually supports the written bug description, filtering inconsistent crowd reports.
DeepPrior (Yu, Wang and Zhang, Prioritize Crowdsourced Test Reports via Deep Screenshot Understanding)	Report text + screenshots + widgets + coordinates + types + intents	Prioritises which crowdsourced reports developers should inspect first, using deep screenshot understanding.
SemCluster (Yu, Fang and Zhang)	Report text + screenshots + screenshot-text binding	Clusters crowdsourced reports using screenshot-text rules, reducing duplicates and reviewer workload.
TIURP (Fang, Yu and Zhang, Enhanced Crowdsourced Test Report Prioritization via Image-and-Text Semantic Understanding and Feature Integration (TIURP))	Report text + screenshots	Prioritises mobile crowdsourced reports using combined image-and-text understanding.
BUGFixChecker (Massenon, Anquetil and Ducasse)	Original report + developer fix claim + before/after UI	Directly a verification dataset — it asks whether a fix really resolved the user's

	screenshots + post-fix reviews; 53 real bug-fix cases	problem (the “fixed but not resolved” gap).
ImageR (Tan, Yadav and Ahmed)	Bug report text + attached images; 34,540 Bugzilla reports across 12 Mozilla products	Studies whether a screenshot helps developers' triage, reproduce, or resolve a bug.
SWE-bench Multimodal (Yang, Zhang and Li)	GitHub issue text + source code + images in problem statements/tests; 617 tasks, 17 JS libraries	Multimodal issue fixing where visual evidence matters — charts, maps, diagrams, and UI components.
VisionDroid / Trident (Liu, Chen and Yu)	Screenshots + GUI text/hierarchy + action history + bug labels; hundreds of reproducible non-crash bugs	Detects non-crash functional bugs (wrong values, counters that fail to update, broken navigation) using visual and logical oracles.
Multi-window GUI defect benchmark (X. Zhang, R. Wang and J. Cui)	Screenshots + widget elements + MLLM prompts + defect labels; 50 real Android apps	Targets split-screen, foldable and multi-window display defects — truncation, reflow — increasingly common on modern devices.
NiCro (Xie, Chen and Yu)	Screenshots + widget detection + visual matching + test cases across devices; 107 cases, 28 apps, 8 devices	Cross-device, cross-platform record-and-replay: a test recorded on one device replays on another despite layout shifts.

Autonomous and cyber-physical system verification.
Not software testing in the GUI sense, but central to verifying AI-enabled perception and autonomy, where sensor fusion and temporal annotation are the whole point (Table 10).

Table 10. Datasets for autonomous and cyber-physical system verification

Dataset	Modalities & scale	Relevance to testing & verification
nuScenes (Caesar, Bankiti and Lang)	6 cameras, 5 radars, 1 LiDAR, GPS/IMU, 3D annotations; 1,000 scenes of 20 s, 23 classes	Verifies that perception software detects, tracks and reasons about traffic participants under real urban conditions.
Waymo Open Dataset (Sun, Kretzschmar and Dotiwalla)	LiDAR + camera + 2D/3D annotations + motion data; 1,150 scenes of 20 s, synchronised and calibrated	Validates safety-critical perception under geographic, lighting and traffic variation.
BDD100K (Yu, Chen and Wang)	Driving video + annotations + GPS/IMU; 100k videos, 10 tasks, weather and time diversity	Tests whether perception generalises across rain, night, highways and dense urban scenes.
Argoverse (Chang, Lambert and Sangkloy)	Multi-sensor data with rich HD maps; 3D tracking and motion-forecasting annotations	Supports map-aware tracking and trajectory-forecasting verification for autonomy stacks.

Matching datasets to testing activities.
Brought together, the landscape can be read as a simple map from a verification activity to the datasets that fit it best (Table 11).

Table 11. Mapping of testing/verification activities to best-fit datasets

Testing / verification activity	Best-fit datasets
Mobile GUI understanding	Rico, MobileViews, ScreenAI, ScreenQA, Screen2Words
UI element detection / visual locators	ScreenSpot, ScreenSpot-Pro, UGround, Widget Captioning
Natural-language test execution	PixelHelp/Seq2Act, AITW, AndroidControl, AMEX, MoTIF
Cross-app mobile workflows	GUI Odyssey (+ CoM), OmniGUI
Web test automation	Mind2Web family, WebArena, VisualWebArena
Enterprise / desktop workflow verification	WorkArena, OSWorld, GUI-360, GUI-World
Dynamic GUI / video testing	GUI-World, OmniGUI, VideoWebArena
Test-report clustering / prioritisation	ReCoDe, DeepPrior, SemCluster, TIURP
Bug-fix & post-fix verification	BUGFixChecker
Multimodal issue fixing / triage	SWE-bench Multimodal, ImageR
Non-crash functional bug detection	VisionDroid/Trident, multi-window defect benchmark
Cross-device GUI test migration	NiCro
Front-end visual verification	WebSRC, Design2Code, Web2Code
Autonomous-system verification	nuScenes, Waymo, BDD100K, Argoverse

Gaps and outlook.
For software testing, the field is steadily moving away from text-only artefacts toward datasets that fuse a requirement, task or bug report with a screenshot or video, a UI hierarchy or DOM, an action trace, logs or code, and — most importantly — an expected behaviour or oracle. That last ingredient is the hardest and rarest: many datasets teach a model to understand a screen or predict an action, but few encode a genuine oracle that says whether the software is behaving correctly. The most testing-specific resources today point the way forward — the crowdsourced-report family (ReCoDe, DeepPrior, SemCluster, TIURP) for triaging noisy human reports; the non-crash functional-bug work (VisionDroid/Trident and the multi-window defect benchmark) for catching wrong-but-not-crashing behaviour; BUGFixChecker for post-fix verification; SWE-bench Multimodal for issue fixing where visuals matter; and NiCro for cross-device replay. For teams building their own corpora the recipe is unchanged: start from these public datasets, add private data and model-generated labels, and spend the scarce human effort on the oracle.

8. Dataset Readiness for LLMs and LMMs

The emergence of LLMs and LMMs has fundamentally changed how data must be collected, processed, and prepared for downstream use. Unlike conventional supervised learning pipelines, where dataset design is primarily guided by label quality and class distribution, modern generative and retrieval-based systems impose a far more nuanced set of requirements on the underlying data. Dataset readiness, in this context, refers not only to the completeness and cleanliness of raw corpora but also to their suitability for specific training paradigms, including supervised fine-tuning, RAG, in-context learning, and multi-agent orchestration. This section provides a systematic treatment of data requirements across these paradigms, covering both LLMs and LMMs, and discusses practical strategies for dataset construction, formatting, and deployment.

8.1. Data Requirements for LLM Fine-Tuning

Fine-tuning a pre-trained LLM involves adapting the model's parameters to a specific task or domain by training on a curated dataset of input-output pairs. The data requirements for this process are distinct from those of general-purpose pre-training in several important ways.

First, instruction-following datasets have become the dominant format for fine-tuning modern LLMs. These datasets consist of triplets that typically include a system prompt, a user instruction, and an expected model response. The quality and diversity of these instructions have a direct bearing on the resulting model's generalisation capability. (Sanh, Webson and Raffel) demonstrated that fine-tuning on multitask prompted datasets, where each task is framed as a natural language instruction, leads to significant improvements in zero-shot performance across unseen tasks. Their work, which introduced the T0 model trained on the P3 dataset, established a key precedent: the breadth and phrasing diversity of instructions matters as much as the volume of data.

Second, data quality filtering is critical for LLM fine-tuning. Noisy or inconsistent training examples can cause the model to learn incorrect generalisations, a problem often referred to as poisoning the instruction-following behaviour. (Wei, Bosma and Zhao) showed that models fine-tuned on high-quality, carefully filtered instruction sets outperform those trained on larger but noisier datasets. This underlines the importance of deduplication, language quality scoring, and toxicity filtering as pre-processing steps before any fine-tuning run.

Third, the format of conversational data must match the model's chat template. Different model families such as LLaMA, Mistral, and GPT-4 use distinct conversation templates, and mismatches between training data format and inference-time prompt structure can degrade performance substantially. (Touvron, Lavril and Izacard) documented the importance of carefully aligning training data format with model architecture expectations, particularly when constructing multi-turn dialogue datasets for chat-oriented fine-tuning.

Fourth, domain adaptation through fine-tuning requires domain-specific corpora that are representative of the target use case. (Gururangan, Marasovic and Swayamdipta) introduced the concept of domain-adaptive pre-training (DAPT), showing that continued pre-training on in-domain data before task-specific fine-tuning leads to consistent performance gains across multiple domains, including biomedical text, computer science, and news. This finding has direct implications for dataset construction: simply collecting more general-purpose data is often insufficient for specialized applications and targeted in-domain corpora must be assembled separately.

8.2. Data Requirements for LMM Fine-Tuning

Large multimodal models extend the LLM paradigm by integrating information from multiple modalities, most commonly image and text, but increasingly also audio, video, and structured tabular data. Fine-tuning LMMs requires datasets that are not only linguistically coherent but also visually grounded and modality aligned.

Image-text alignment is the fundamental challenge in constructing LMM training data. Each image in the dataset must be paired with text that meaningfully describes or references its visual content, rather than simply co-occurring with it. (Liu, Li and Wu) introduced the LLaVA model and its associated training dataset, which was constructed by using GPT-4 to generate instruction-following data grounded in image captions and bounding box annotations. This approach highlighted a key insight: automatically generated visual instructions, when produced by a sufficiently capable model, can match or exceed the quality of manually annotated datasets for certain tasks.

Beyond image-caption pairs, fine-tuning LMMs for complex reasoning tasks requires visual question answering (VQA) data, scene graph annotations, and interleaved image-text sequences. (Alayrac, Donahue and Luc) demonstrated through the Flamingo model that training on large quantities of interleaved image-text data, where images and text appear in natural co-occurrence rather than as paired examples, significantly improves few-shot performance on multimodal tasks. The diversity of

visual contexts encountered during training plays a similar role to instruction diversity in LLM fine-tuning.
For video-language tasks, temporal alignment between visual frames and textual descriptions introduces additional complexity. Datasets must capture not only what appears in each frame but also the temporal relationships between events. (Zellers, Lu and Hessel) addressed this in the context of video understanding by constructing datasets that require models to reason about causal and temporal event structures, a requirement that demands careful annotation protocols.
A practical challenge in LMM dataset preparation is the resolution and quality of visual inputs. Low-resolution images or images with significant compression artifacts can degrade model performance, particularly for tasks that require fine-grained visual reasoning such as reading text in images or identifying small objects. (Dai, Li and Li) reported that InstructBLIP, an instruction-tuned multimodal model, showed significant sensitivity to image preprocessing choices, emphasising the need for consistent and high-quality visual data pipelines.

8.3. Retrieval-Augmented Generation (RAG) Data Requirements

Retrieval-Augmented Generation (RAG) is a paradigm that augments LLM inference by retrieving relevant documents from an external knowledge base and conditioning the model's generation on the retrieved content. Unlike fine-tuning, RAG does not modify model parameters; instead, it relies on the quality and organisation of an external document corpus to extend the model's effective knowledge boundary.
The data requirements for RAG systems are fundamentally different from those for fine-tuning. The primary concern shifts from training example quality to corpus organisation, indexing fidelity, and retrieval relevance. (Lewis, Perez and Piktus) introduced the RAG framework and demonstrated that retrieval quality has a more pronounced effect on downstream task performance than the size of the document corpus, a finding that has important implications for how RAG data pipelines should be designed.
Document chunking is one of the most critical decisions in RAG data preparation. A chunk is a segment of a document that is stored and retrieved as a unit. If chunks are too large, the retrieved context may contain irrelevant information that distracts the model; if chunks are too small, individual chunks may lack sufficient context for the model to generate a coherent response. (Shi, Chen and Misra) demonstrated through controlled experiments that the granularity of retrieved context significantly affects how well LLMs can integrate external information, with paragraph-level chunks generally outperforming both sentence-level and full-document retrieval for most question-answering tasks.
Beyond raw chunking, the structure of the retrieved document matters. Tables, figures, and structured lists are often poorly represented in plain-text chunks, leading to retrieval pipelines that lose structured information. (Borgeaud, Mensch and Hoffmann) noted in the RETRO paper that the format and preprocessing of the retrieval corpus substantially affect model performance, and that some structured content types require specialised extraction procedures before they can be effectively used in RAG pipelines.

8.4. Prompt-Aware Dataset Design

Prompt-aware dataset design refers to the practice of constructing training and evaluation datasets with explicit attention to how the data will be presented to the model at inference time. As LLMs have become increasingly sensitive to prompt formatting, the field has recognised that dataset design cannot be fully decoupled from prompt engineering.
The seminal work by (Brown, Mann and Ryder) on GPT-3 introduced the concept of in-context learning and demonstrated that model performance varies substantially depending on the phrasing and structure of prompts, even when the underlying task remains the same. This observation

motivated a new generation of dataset design practices in which examples are constructed to be compatible with specific prompt templates rather than as free-form input-output pairs.
(Mishra, Khashabi and Baral) introduced the CROSS-TASK GENERALISATION benchmark and showed that models trained on datasets with explicitly defined prompt templates generalise better to new tasks when the test-time prompts follow a similar structure. This suggests that dataset creators should adopt consistent prompt templates across examples and document these templates as part of the dataset release, so that users can align their inference prompts accordingly.
Template-based dataset construction also interacts with the choice of delimiter tokens, special markers, and role labels in chat-formatted datasets. (Voronov, Wolf and Ryabinin) showed that seemingly minor formatting decisions, such as whether to include a newline after the instruction or how to delimit the response, can result in non-trivial performance differences on downstream benchmarks. These findings argue for careful documentation of dataset formatting conventions alongside the data itself.

8.5. Chunking and Indexing Strategies

The effectiveness of a RAG system depends critically on how the document corpus is divided into retrievable units and how these units are indexed for efficient search. Chunking and indexing strategies form the data infrastructure layer that determines both retrieval accuracy and computational efficiency.
Several chunking strategies have been proposed in the literature, ranging from fixed-size character or token windows to semantically aware splits based on sentence boundaries or topic shifts. (Gao, Xiong and Gao) conducted a systematic evaluation of these strategies and found that semantic chunking, where chunks are defined by topic coherence rather than fixed length, consistently produces better retrieval performance across diverse query types, at the cost of increased complexity in the preprocessing pipeline.
Hierarchical indexing is an emerging strategy that maintains multiple levels of document representation simultaneously. In a hierarchical index, the system stores both high-level document summaries and fine-grained passage chunks, allowing the retriever to first identify relevant documents and then locate the most relevant passages within them. This approach was shown to reduce retrieval noise while preserving recall in the work of (Sarthi, Abdullah and Tuli), who introduced the RAPTOR method for recursive abstractive processing and tree-organised retrieval.
Dense retrieval, which uses neural embeddings to represent both queries and document chunks in a shared vector space, has become the dominant indexing paradigm for modern RAG systems. (Karpukhin, Oguz and Min) introduced Dense Passage Retrieval (DPR) and demonstrated that bi-encoder models trained on question-passage pairs substantially outperform sparse retrieval methods such as BM25 on open-domain question answering tasks. The quality of the dense index depends not only on the embedding model but also on how the document chunks are prepared before encoding.
Hybrid indexing, which combines dense semantic search with sparse keyword-based retrieval, has been shown to offer the best of both approaches for real-world applications. (Robertson and Zaragoza) established the strong baseline performance of BM25 for keyword-based retrieval, and subsequent work by (Lin, Ma and Lin) demonstrated that combining BM25 with dense retrieval using reciprocal rank fusion consistently improves retrieval quality over either method alone. This hybrid approach has become a practical standard for production RAG systems.

8.6. Dataset Preparation for Few-Shot and In-Context Learning

Few-shot and in-context learning represent a paradigm where the model is provided with a small number of labelled examples as part of the input prompt, rather than being fine-tuned on a large,

annotated dataset. Preparing datasets for this paradigm requires attention to example selection, ordering, and diversity.

The key insight from (Brown, Mann and Ryder) is that LLMs can learn from examples presented in the prompt without gradient updates, a capability that fundamentally changes the requirements for dataset construction. Rather than building large training corpora, practitioners must instead curate high-quality example pools from which demonstrations can be selected at inference time.

Example selection has been shown to have a large effect on few-shot performance. (Liu, Shen and Zhang) introduced the concept of dynamic example selection, where demonstrations are chosen based on their semantic similarity to the current test query rather than randomly sampled from a fixed set. Their approach, using retrieval-based example selection, substantially improved few-shot performance across multiple classification and generation tasks.

The ordering of examples within a few-shot prompt also matters. (Lu, Bartolo and Moore) systematically studied the effect of demonstration order and found that performance can vary by several percentage points depending on which examples appear first and which appear last in the prompt context. These findings suggest that dataset curators should provide guidance on example ordering alongside their datasets, particularly when the examples are intended for few-shot use.

Diversity of examples is a further consideration. If all few-shot examples belong to the same semantic cluster or exhibit similar surface patterns, the model may fail to generalise across the range of inputs it encounters at inference time. (Hongjin, Kasai and Wu) proposed a diversity-driven example selection strategy and demonstrated that maximising coverage of the input space in the selected demonstrations leads to more robust few-shot performance.

8.7. Dataset Preparation for Multi-Agent Systems

Multi-agent systems represent a more complex deployment paradigm in which multiple LLM-based agents collaborate to complete tasks, each with access to specific tools, memory stores, and communication channels. Preparing datasets for multi-agent systems requires extending conventional dataset design to account for inter-agent communication, tool use, and sequential decision-making.

The ReAct framework proposed by (Yao, Zhao and Yu) introduced a data format that interleaves reasoning traces with action invocations, enabling models to learn to use external tools as part of their response generation. Datasets formatted in the ReAct style include not just input-output pairs but complete trajectories of thought and action, which must be collected either through human demonstration or by prompting a strong model to generate them.

Toolformer, introduced by (Schick, Dwivedi-Yu and Dessì), took a self-supervised approach to generating tool-use training data. By prompting a model to generate candidate API calls and filtering those that improve its own performance, Toolformer demonstrated that it is possible to construct high-quality tool-use datasets without extensive human annotation. This method is particularly relevant for multi-agent systems, where the action space may be large and difficult to annotate manually.

For systems involving multiple agents collaborating on complex tasks, datasets must capture the division of labour between agents, the passing of partial results, and the aggregation of individual contributions. (Park, O'Brien and Cai) introduced a simulated social environment where multiple LLM-based agents generate observations about their environment, form intentions, and communicate with one another. The resulting interaction logs constitute a novel dataset format that captures the dynamics of multi-agent collaboration.

Memory management is a further dimension of dataset preparation for multi-agent systems. Long-horizon tasks require agents to maintain and update beliefs about the state of the world across many interaction steps. Datasets designed for multi-agent learning must therefore include not only the raw interaction history but also structured representations of agent state, task progress, and

shared context. Preparing such datasets requires careful design of the state representation format, the granularity of memory updates, and the mechanisms for sharing information across agents.

9. Sustainability and Resource-Efficient Data Practices

As large-scale AI systems have grown in scope, the environmental and economic costs associated with data collection, processing, and storage have attracted increasing scrutiny from both the research community and the public. The creation of datasets for training LLMs and LMMs is not a cost-free activity: it involves substantial expenditure in human annotation labour, computational resources for preprocessing and filtering, and energy consumption for data storage and transfer. Moreover, the rapid turnover of dataset versions and the tendency to collect new data rather than reuse existing resources compounds these costs at a systemic level. This section examines the sustainability dimensions of data practices for large-scale AI systems, covering the cost trade-offs between collection and reuse, techniques for data reduction and compression, strategies for selective modality usage, principles of dataset lifecycle management, and the carbon footprint implications of dataset creation.

9.1. Cost of Data Collection Versus Data Reuse

The decision to collect new training data or to reuse existing datasets is one of the most consequential choices in the development of large-scale AI systems. From a purely economic perspective, the cost of collecting high-quality annotated data from scratch is substantial, involving recruitment and compensation of human annotators, quality assurance processes, and infrastructure for data collection platforms. Reusing existing datasets avoids many of these costs but introduces its own challenges related to compatibility, license compliance, and dataset drift.

(Bommasani, Hudson and Adeli) noted in their foundational survey of foundation models that the datasets used to train these systems represent an enormous and largely invisible infrastructure investment. The Common Crawl corpus, for example, is the product of years of sustained web crawling and forms the backbone of many LLM pre-training pipelines. While freely available, the costs of cleaning, filtering, and deduplicating such corpora before use are non-trivial and are often underreported in published work.

The economic argument for data reuse is compelling when the existing data is of sufficient quality and domain relevance. (Raffel, Shazeer and Roberts) demonstrated in the context of the C4 dataset, a cleaned version of Common Crawl, that aggressive filtering and deduplication of an existing web corpus is substantially more cost-effective than collecting new text data from scratch, while producing a dataset suitable for pre-training competitive language models. This work established data cleaning and reuse as a viable primary strategy for large-scale dataset construction.

However, the cost calculus shifts when existing datasets are insufficiently representative of the target domain or task. In specialised fields such as medicine, law, and scientific research, publicly available corpora may be sparse, outdated, or restricted by access controls, making targeted data collection a practical necessity despite its higher cost. (Lehman, Hernandez and Mahajan) illustrated this challenge in the biomedical domain, where the gap between general-purpose corpora and clinical text necessitated the collection of domain-specific datasets to achieve acceptable performance on clinical NLP tasks.

Dataset licensing is an often overlooked but significant factor in the reuse equation. Many popular datasets carry restrictive licenses that prohibit commercial use or redistribution, effectively limiting their reuse to academic contexts. (Lhoest, del Moral and Jernite) documented this issue in the context of the Hugging Face Datasets library, noting that license ambiguity or restriction affects a substantial proportion of publicly available datasets and creates legal uncertainty for downstream

users. Clear licensing practices and the production of openly licensed datasets are therefore important contributions to a more sustainable data ecosystem.

9.2. Data Reduction and Compression

The scale of datasets used in modern AI training pipelines has grown to the point where storage and transmission costs represent a meaningful fraction of total project expenditure. Data reduction and compression techniques address this challenge by decreasing the footprint of datasets without proportionally degrading their utility for model training.

Dataset pruning, the removal of redundant or low-utility training examples, is one of the most widely studied data reduction techniques. (Sorscher, Geirhos and Shekhar) demonstrated through theoretical and empirical analysis that intelligently pruning a dataset can improve model training efficiency, and that at small data scales, a well-pruned subset can outperform the full dataset. Their work introduced a geometry-based pruning criterion that identifies and removes examples that are easy for the model and contributes little gradient signal during training.

Data distillation, sometimes called dataset condensation, takes a more aggressive approach by replacing the original dataset with a small synthetic dataset that is optimised to produce the same model behaviour as training on the full data. (Zhao, Mopuri and Bilen) introduced dataset condensation with gradient matching and showed that models trained on distilled datasets of just a few images per class can achieve performance competitive with models trained on the full dataset for certain classification tasks. While the computational cost of producing a distilled dataset can be high, the resulting dataset is far smaller and can be stored and transferred much more efficiently.

Deduplication is a form of data reduction that specifically targets redundant examples, near-identical data points that add little new information to the training corpus. (Lee, Ippolito and Nystrom) demonstrated that removing near-duplicate examples from large pre-training corpora substantially reduces memorisation, improves generalisation, and decreases the size of the dataset without harming downstream task performance. Their MinHash-based deduplication method has since become a standard pre-processing step in large-scale LLM data pipelines.

Compression at the storage level, using formats such as Parquet, LMDB, or custom binary encodings, reduces the physical storage requirements of large datasets. (Lhoest, del Moral and Jernite) noted that the adoption of efficient serialisation formats in the Hugging Face Datasets library reduced memory consumption during dataset loading by an order of magnitude compared to naive CSV or JSON storage, enabling researchers with limited hardware resources to work with large datasets that would otherwise be inaccessible.

9.3. Selective Modality Usage

Multimodal AI systems require data from multiple input modalities (text, image, audio, video, and structured data) and the collection, storage, and processing of all modalities simultaneously is resource intensive. Selective modality usage refers to the practice of carefully choosing which modalities to include in a dataset based on their contribution to task performance and their associated resource costs.

Not all modalities contribute equally to model performance on a given task. (Akbari, Yuan and Qian) showed in the context of video-language understanding that the relative importance of visual and audio modalities varies significantly across tasks, and that models trained with access to all modalities do not always outperform those trained on a carefully selected subset. This finding suggests that uncritical inclusion of all available modalities can increase data collection and processing costs without corresponding performance benefits.

The resolution and quality of visual modalities represent a particularly important axis for resource optimisation. High-resolution images or video consume significantly more storage and processing resources than their lower-resolution counterparts. (Dosovitskiy, Beyer and Kolesnikov) showed in

the Vision Transformer (ViT) paper that the optimal image resolution for a given task depends on the complexity of the visual information required, and that using unnecessarily high resolutions increases computational costs without improving accuracy on many tasks. These findings support the practice of selecting the minimum resolution sufficient for the task at hand.

Temporal modalities such as video and audio require additional consideration because they encode information along a time axis. Subsampling video at a lower frame rate or using shorter audio clips can substantially reduce data volume without necessarily degrading model performance, provided that the subsampling strategy preserves the informative portions of the signal. (Feichtenhofer, Fan and Malik) demonstrated in the context of video action recognition that temporal subsampling strategies have a significant effect on both model performance and training efficiency, and that optimal sampling rates are task dependent.

Modality fusion choices also affect the resource requirements of downstream model training. Late fusion strategies, where each modality is processed independently before combination, allow for greater flexibility in selectively using or omitting modalities without redesigning the entire model architecture. This contrasts with early fusion approaches that tightly couple modalities from the input stage, making selective modality usage more difficult after the fact. (Baltrusaitis, Ahuja and Morency) provided a comprehensive taxonomy of multimodal fusion strategies and discussed the trade-offs between flexibility and representational capacity that bear on resource-efficient dataset design.

9.4. Dataset Reuse and Lifecycle Management

Beyond the initial decision of whether to collect or reuse data, sustainable data practices require attention to the full lifecycle of a dataset, from creation and release through to maintenance, versioning, and eventual deprecation. Poor lifecycle management leads to datasets that become outdated, misused, or unavailable to the research community.

Dataset versioning is a fundamental component of lifecycle management. As datasets are updated, corrected, or extended, maintaining clear version histories allows users to reproduce results from prior publications and to understand how the dataset has evolved. (Gebru, Morgenstern and Vecchione) introduced the concept of datasheets for datasets, structured documentation that records the dataset's intended uses, collection methodology, composition, and recommended updates. The adoption of datasheets has been shown to improve reproducibility and to reduce misuse of datasets in contexts for which they were not designed.

The concept of dataset drift refers to the decreasing relevance of a dataset over time as the world it was meant to represent changes. Text corpora drawn from web crawls become outdated as language evolves and new topics emerge. (Lazaridou, Kuncoro and Gribovskaya) studied temporal concept drift in language models and showed that models trained on older data exhibit systematic degradation on questions about recent events, a problem that is exacerbated when datasets are reused without updates. Sustainable dataset practices must therefore include mechanisms for identifying when a dataset has drifted beyond its useful life and requires refreshing.

Data provenance tracking, recording where each data point came from, how it was processed, and what transformations were applied, is an important tool for lifecycle management. Provenance information enables auditing of training data for biases, errors, and license violations, and supports the removal of specific data points in response to data subject requests under privacy regulations such as the GDPR. (Hutchinson, Smart and Hanna) argued that treating dataset development as a continuous and accountable process, rather than a one-time activity, is essential for long-term sustainability.

Benchmark saturation is a lifecycle issue that affects evaluation datasets rather than training corpora. As models are trained on data that closely resembles popular benchmarks, the discriminative power of those benchmarks decreases. (Raji, Bender and Paullada) documented

benchmark saturation across multiple NLP tasks and argued for the development of dynamic evaluation frameworks where test sets are periodically refreshed to prevent overfitting at the ecosystem level. This requires treating evaluation datasets as living resources that require ongoing curation and management.

9.5. Carbon Footprint of Dataset Creation

The carbon footprint of AI systems has received increasing attention as the energy consumption of large-scale model training has become widely documented. However, the environmental impact of dataset creation itself, including the computational resources used for web crawling, preprocessing, filtering, and storing large corpora, is less frequently discussed and represents an important dimension of sustainable AI development.

(Strubell, Ganesh and McCallum) provided one of the earliest systematic analyses of the carbon cost of training large NLP models and found that the energy consumption of training a single model can be comparable to the lifetime energy consumption of several automobiles. While this analysis focused on model training, the data preparation pipeline, which may involve running multiple large-scale preprocessing jobs over hundreds of terabytes of raw data, contributes a non-trivial share of total project energy consumption that was not fully accounted for in their estimates.

(Patterson, Gonzalez and Le) provided an updated analysis of the energy and carbon footprint of deep learning, noting that hardware efficiency, data centre energy sources, and geographic location of computational infrastructure are among the most important determinants of carbon impact. Their findings suggest that practitioners can substantially reduce the carbon footprint of dataset creation by choosing cloud infrastructure powered by renewable energy and by scheduling intensive preprocessing jobs during periods of low-carbon electricity availability.

(Lottick, Susai and Friedler) developed tools for tracking the carbon footprint of machine learning experiments and demonstrated that the variance in carbon emissions across equivalent computational tasks is high, primarily due to differences in the carbon intensity of the electricity grid at the time and location of computation. This underscores the importance of making carbon accounting an integrated part of the data pipeline rather than an afterthought.

(Schwartz, Dodge and Smith) introduced the concept of Green AI and argued that the field's focus on benchmark performance has created incentives that systematically neglect computational efficiency and environmental impact. They called for the inclusion of computational cost and carbon impact as first-class metrics in research publications, a practice that has been partially adopted by some conferences through efficiency tracks and reporting requirements. Applying this principle to dataset creation means reporting not only the scale and characteristics of a dataset but also the resources consumed in its production.

(Lacoste, Lottick and Goyal-Kamal) introduced the Machine Learning Emissions Calculator, a practical tool that allows researchers to estimate the carbon emissions of their computational workloads based on hardware type, duration, and geographic location. The availability of such tools makes carbon accounting feasible for individual researchers and teams, and their adoption in data pipeline planning can support more informed decisions about the trade-offs between data scale and environmental impact.

Reproducibility considerations intersect with carbon footprint concerns in an important way. When dataset preprocessing pipelines are not well documented, downstream users may repeat expensive preprocessing steps unnecessarily, compounding the carbon cost of the original data creation. Sustainable data practices therefore include thorough documentation of preprocessing procedures, the release of pre-processed dataset versions alongside raw data, and the use of checkpointing to avoid redundant computation in iterative processing pipelines.

10. Summary of Gaps and CLEAR Positioning

This section consolidates the key gaps identified in the preceding state-of-the-art (SotA) review and positions CLEAR's WP2 guidelines accordingly. The intent is not to restate the technical literature, but to translate the reviewed evidence into an actionable, project-internal reference framework that will steer dataset creation, collection, and fusion activities across Tasks 2.2–2.4 and ensure downstream suitability for WP3 (model construction and fine-tuning) and WP4 (multimodal pipeline development).

10.1. Summary of gaps identified in the SotA

Across the reviewed bodies of work on multimodal dataset creation, synthetic data generation, fusion strategies, and dataset readiness for LLMs/LMMs, the following gaps are particularly salient for industrial deployment contexts such as those addressed by CLEAR.

1. Limited availability of industrial-grade multimodal datasets: Public datasets provide breadth across modalities, but they rarely capture the combination of operational constraints, heterogeneous sensing environments, and domain-specific semantics that characterise industrial settings. Many benchmark datasets are either consumer-oriented, generated under comparatively controlled conditions, or insufficiently documented for traceability and reuse in regulated environments. As a result, their direct applicability to industrial use cases is limited, especially where multimodal alignment must be robust under real-world noise, legacy formats, and non-standard metadata practices.
2. Insufficiently operationalised guidance for multimodal dataset construction: The SotA provides numerous modality-specific recommendations (e.g., for image/video synthesis or time-series generation) yet offers comparatively less concrete guidance on how to build *end-to-end* industrial datasets that remain coherent across modalities and across lifecycle phases (collection, curation, annotation, versioning, update, and reuse). This creates a practical gap between conceptual best practices and what consortia can implement consistently across partners and domains.
3. Persistent challenges in multimodal alignment and fusion under heterogeneity: While fusion taxonomies (early, intermediate, late, hybrid) are well established, the literature does not converge on decision criteria that are sufficiently explicit for industrial project execution. In practice, industrial modalities differ substantially in sampling rate, semantic granularity, reliability, and availability. Moreover, industrial data is frequently asynchronous and partially missing and may be subject to distribution shifts due to changing equipment, processes, or documentation practices. These characteristics make fusion strategy selection highly context-dependent, yet SotA guidance often remains generic.
4. Data scarcity, long-tail events, and the limits of naive synthetic augmentation: Industrial datasets commonly exhibit long-tail distributions, with rare but safety- or cost-critical events underrepresented. The SotA indicates strong momentum toward synthetic data generation and hybrid data strategies; however, it also highlights the associated risks, including artefact learning, biased augmentation, and degradation under recursive synthetic training. Validation protocols for synthetic multimodal data, particularly those capturing cross-modal consistency rather than unimodal plausibility, remain immature in many applied settings.
5. Confidentiality, ownership, and constrained data sharing as first-order design constraints: Industrial data handling is structurally shaped by confidentiality, intellectual property protection, privacy requirements, and contractual limitations on data exchange. While secure data handling methods exist, many SotA dataset practices implicitly assume open data flows or permissive sharing regimes. This mismatch leads to practical barriers for cross-partner dataset creation and complicates benchmarking, reproducibility, and the development of shared multimodal resources.

6. Dataset readiness for LLM/LMM pipelines is unevenly addressed: The shift from conventional supervised learning to LLM/LMM-centric paradigms (fine-tuning, RAG, instruction tuning, in-context learning, and tool-using/multi-agent workflows) imposes new constraints on dataset structure, formatting, and documentation. The SotA acknowledges these paradigms, but industrial dataset practices often lag behind: corpora are frequently not prepared with prompt-awareness, consistent chat templates, retrievable chunk structure, or modality-aligned instruction formats in mind. This creates friction when translating curated industrial data into usable training and evaluation assets for WP3/WP4.

10.2. CLEAR positioning: how WP2 addresses the identified gaps

Against these gaps, CLEAR positions WP2 as the methodological backbone that enables coherent, industrially grounded multimodal datasets and fusion-ready data assets across the consortium.
CLEAR's approach is characterised by the following principles:

- Domain-agnostic but industrially constrained by design: WP2 guidelines are constructed to remain applicable across CLEAR's diverse industrial sectors, while explicitly incorporating the constraints that typically differentiate industrial data work from academic dataset development (confidentiality, access control, heterogeneous legacy systems, and limited labelling capacity). The aim is a shared dataset methodology that is reusable across use cases without assuming uniform tooling, formats, or data governance structures.
- Fusion-oriented dataset design rather than post-hoc modality linking: CLEAR treats multimodal fusion requirements as a *dataset design input*, not only as a modelling-stage decision. WP2 therefore emphasises modality alignment practices, metadata harmonisation, and the documentation of temporal/semantic relationships between modalities as prerequisites for meaningful early/intermediate/hybrid fusion experiments later in the project.
- Hybrid data strategies coupled with validation and traceability: CLEAR adopts hybrid data strategies to mitigate industrial scarcity and long-tail effects, but positions synthetic data generation as a controlled complement rather than a replacement for real data. WP2 guidelines therefore foreground validation, artefact risk management, and provenance tracking (including explicit recording of which data is real, synthetic, transformed, or anonymised) to support reliable training and evaluation.
- Confidentiality-aware dataset workflows enabling consortium-level progress: WP2 defines practical routes to enable meaningful cross-partner development without requiring unrestricted data sharing. This includes controlled access principles, anonymisation/pseudonymisation guidance where feasible, and the use of representative or synthetic substitutes when raw data cannot leave organisational boundaries. The objective is to ensure that confidentiality constraints do not translate into methodological fragmentation across use cases.
- Explicit "dataset readiness" for LLMs and LMMs: CLEAR positions dataset readiness as a core WP2 outcome: datasets are not considered complete when they are merely collected and cleaned, but when they are *deployable* for the targeted model-development paradigms. This includes guidance on instruction formats, retrieval-oriented structuring, consistent documentation, and the definition of evaluation subsets aligned with expected downstream tasks.

10.3. Implications for WP2 guidelines and downstream work

The SotA review indicates that the primary risk for industrial multimodal AI is not the absence of algorithms, but the absence of consistent, validated, and shareable data foundations that permit

those algorithms to be applied reliably under industrial constraints. CLEAR therefore operationalises the SotA through WP2 guidelines that:

- standardise how multimodal data is collected, documented, and versioned across partners;
- define alignment and fusion-relevant metadata as first-class dataset components;
- enable hybrid (real + synthetic) dataset construction with explicit validation expectations;
- embed confidentiality-aware practices as default, not as an afterthought; and
- ensure that resulting datasets are structured for the LLM/LMM workflows envisaged in WP3 and WP4.

In this way, the WP2 guidelines provide the transition from “what is known” in the SotA to “what will be implemented” in CLEAR, ensuring coherence of data practices across the consortium and providing an auditable basis for subsequent model development and evaluation.

*Table 4: **Gap-to-guideline mapping (WP2).** The "Identified gap" column summarises what the SoTA does not sufficiently resolve for industrial multimodal data work; The "WP2 guideline focus" column states how CLEAR turns these gaps into implementable dataset rules and decisions; The "Expected artefacts" column lists concrete outputs that can be produced and maintained throughout WP2 to ensure downstream usability in WP3/WP4.*

Identified gap (from SotA)	Consequence for industrial deployment	WP2 guideline in CLEAR	Expected WP2 artefact(s) / output(s)
Limited availability of industrial-grade multimodal datasets	Public datasets are not representative of industrial heterogeneity, noise, legacy formats, and governance constraints	Establish a common dataset blueprint for *industrial* multimodal data across domains and partners	Dataset blueprint per use case; agreed minimum dataset specification (modalities, metadata, governance); initial dataset inventory and data source register
Insufficiently operationalised end-to-end guidance	Partners may implement inconsistent collection/annotation practices, reducing reuse and comparability	Provide stepwise guidance for collection, curation, annotation, documentation, versioning, and reuse	WP2 dataset creation checklist; documentation template (incl. datasheet-style fields); versioning and change-log conventions
Persistent multimodal alignment and fusion challenges	Fusion experiments become brittle due to missing alignment signals and unclear assumptions	Treat alignment as a dataset property: define required linkage keys, temporal windows, semantic anchors, and synchronisation assumptions	Alignment specification per dataset (time sync policy, entity IDs, reference frames); modality linking rules; defined "alignment confidence" metadata fields
Data scarcity and long-tail event under-representation	Models underperform on rare but critical cases; biased learning	Adopt hybrid strategies combining real data with controlled synthetic augmentation, targeted at coverage gaps	Synthetic data generation plan per modality; long-tail augmentation plan; defined coverage targets (events/classes/conditions)
Risks of naive synthetic augmentation (artefacts, bias, model collapse)	Synthetic data can degrade robustness or distort distributions	Define validation and acceptance criteria for synthetic data, including cross-modal consistency checks where relevant	Synthetic data validation protocol; acceptance thresholds (utility/diversity/privacy); audit trail capturing generator, prompts/parameters, and intended use
Confidentiality, IP, privacy, constrained sharing	Cross-partner dataset work is blocked or fragmented; limited benchmarking	Make confidentiality-aware workflows the default: controlled access, anonymisation where feasible, and use of representative/synthetic substitutes	Data governance and access matrix; anonymisation/pseudonymisation guidance; secure sharing workflow; "shareable subset" definition per dataset

Identified gap (from SotA)	Consequence for industrial deployment	WP2 guideline in CLEAR	Expected WP2 artefact(s) / output(s)
Dataset readiness for LLM/LMM pipelines unevenly addressed	Collected data is not readily usable for fine-tuning, RAG, or multimodal instruction formats	Define readiness criteria for fine-tuning, RAG, and evaluation subsets; enforce consistent formatting and documentation	Prepared fine-tuning formats (instruction/chat templates); RAG-ready corpus structure (chunking, metadata); evaluation splits aligned with tasks
Sustainability and lifecycle management often under-specified	High overhead from reprocessing, duplicated storage, unmanaged drift	Promote reuse, pruning/deduplication, efficient storage formats, and lifecycle/refresh planning	Dataset lifecycle plan; deduplication strategy; storage format guidance; refresh triggers (drift indicators)

References

Ahsen, M, E, M, U, S Ayvaci and S Raghunathan. “When algorithmic predictions use human-generated data: A bias-aware classification algorithm for breast cancer diagnosis.” *Information Systems Research* 30.1 (2019): 97–116.

Akbari, H, et al. “VATT: Transformers for multimodal self-supervised learning from raw video, audio and text. .” *Advances in Neural Information Processing Systems (NeurIPS)* 34 (2021).

Alam, N, B, et al. “Challenges and standardisation strategies for sensor-based data collection for digital phenotyping.” *Communications Medicine* 5.1 (2025): 360.

Alam, T, M, et al. “An investigation of credit card default prediction in imbalanced datasets.” *IEEE Access* 8 (2020): 201173–201198.

Alayrac, J, B, et al. “Flamingo: a visual language model for few-shot learning.” *Advances in neural information processing systems* 35 (2022): 23716-23736.

Alexandrov, A, et al. “Gluonts: Probabilistic and neural time series modeling in python.” *Journal of Machine Learning Research* 21.116 (2020): 1-6.

Alicea, M, A, A. “Enhancing Social Media User Privacy: Applying Machine Learning-Based Anonymization and Pseudonymization Techniques.” *ProQuest Dissertations & Theses Global* (2025).

Al-Kabbany, A and E Kassem. “Synthesizing the Expert: A Validated Multimodal Dataset for Trustworthy AI-Assisted Swimming Coaching.” *arXiv preprint* (2026): arXiv:2605.12799. <https://arxiv.org/abs/2605.12799>.

Alrayzah, A. “Tri-MCA fusion: cross-modal attention and dynamic gating for multimodal sentiment analysis.” *Scientific Reports* (2026).

Ango, R, et al. “Fraud detection in banking using the Kaggle credit card dataset and XGBoost model.” *In 2024 International Conference on IoT Based Control Networks and Intelligent Systems* (2024).

Arpaci, I. “A multianalytical SEM-ANN approach to investigate the social sustainability of AI chatbots based on cybersecurity and Protection Motivation Theory. .” *IEEE Transactions on Engineering Management* 71 (2024): 1714–1725.

Arshad, M, U, et al. “A privacy mechanism for access controlled graph data.” *IEEE Transactions on Dependable and Secure Computing* 16.5 (2017): 819-832.

Baechler, Géraldine, et al. “ScreenAI: A Vision-Language Model for UI and Infographics Understanding.” *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI)*. International Joint Conferences on Artificial Intelligence Organization (IJCAI), 2024.

Baltrusaitis, T, C Ahuja and L, P Morency. “Multimodal machine learning: A survey and taxonomy.” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 41.2 (2019): 423–443.

Bernard. “Ethical Use of AI.” *Security Technology & Design* 35.3 (2025): 14–15.

Bienzeisler, J, et al. “Implementation report on pioneering federated data access for the German National Emergency Department Data Registry.” *NPJ digital medicine* 8.1 (2025): 94.

Biswal, S, P and M, S Kulkarni. “Implications of GDPR on emerging technologies in healthcare.” *Cardiometry* (2022).

Bodhe, S, et al. “Diagnostic Fail Data Minimization Using an N-Cover Algorithm.” *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* 24.3 (2015): 1198–1202.

Bommasani, R, et al. “On the opportunities and risks of foundation models.” *arXiv preprint* (2021): arXiv:2108.07258.

Borgeaud, S, et al. “Improving language models by retrieving from trillions of tokens.” *International conference on machine learning* PMLR.June (2022): pp. 2206-2240.

Branton, R, et al. “Using Metadata Standards to Achieve Data Interoperability: Proceedings of a Technical Symposium and Workshop.” *Proceedings Series* (2006).

Brown, T, et al. “Language models are few-shot learners.” *Advances in neural information processing systems* 33 (2020): 1877-1901.

Burns, Andrea, et al. “A Dataset for Interactive Vision-Language Navigation with Unknown Command Feasibility.” *European Conference on Computer Vision (ECCV)*. Springer, 2022. 312–328.

Cabon, Yohann, Naila Murray and Martin Humenberger. “Virtual KITTI 2.” *arXiv* (2020): 2001.10773.

Caesar, H, et al. “nuscenes: A multimodal dataset for autonomous driving.” *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020): pp. 11621-11631).

Cardin, O, et al. “Future industrial systems: Best practices of the intelligent manufacturing and services systems (IMS2) French research group.” *IEEE Transactions on Industrial Informatics* 13.2 (2017): 704–713.

Carlesso, A, et al. “Data quality assessment of aggregated LCI datasets: A case study on fossil-based and bio-based plastic food packaging.” *Journal of Industrial Ecology* 28.6 (2024): 1900–1911.

Ceravolo, P, et al. “HH4AI: A methodological framework for AI human rights impact assessment under the EU AI Act.” *arXiv* (2025).

Chai, Yuxuan, et al. “AMEX: Android Multi-Annotation Expo Dataset for Mobile GUI Agents.” 2024. <https://arxiv.org/abs/2407.17490>.

Chang, Ming-Fang, et al. “Argoverse: 3D Tracking and Forecasting with Rich Maps.” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019. 8748–8757.

Chen, Dong, et al. “GUI-World: A Dataset for GUI-Oriented Multimodal LLM-Based Agents.” 2024. <https://arxiv.org/abs/2406.10819>.

Chen, F, et al. “Hierarchical spatio-temporal graph network for risk prediction.” *Scientific Reports* 16.1 (2025): 3853.

Chen, Guikun and Wenguan Wang. “A Survey on 3D Gaussian Splatting.” *arXiv* (2026): 2401.03890v9.

Chen, L, et al. “Sharegpt4v: Improving large multi-modal models with better captions.” *In European Conference on Computer Vision* (2024): pp. 370-387.

Chen, S, et al. “HAIPipe: Combining human-generated and machine-generated pipelines for data preparation.” *Proceedings of the ACM on Management of Data* 1.1 (2023).

Chen, Xiang, et al. “WebSRC: A Dataset for Web-Based Structural Reading Comprehension.” *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2024. 4173–4185.

Cheng, Ke, et al. “SeeClick: Harnessing GUI Grounding for Advanced Visual GUI Agents.” *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics (ACL), 2024.

Chern, E, et al. “Thinking with Generated Images.” *arXiv preprint* (2025): arXiv:2505.22525. <https://arxiv.org/abs/2505.22525>.

Conti, M, et al. “Mobile Networks for Biometric Data Analysis.” *Lecture Notes in Electrical Engineering* 392 (2016).

Dabbaghchian, I, et al. “Optimizing bridge modal property estimation via quality-driven crowdsourced smartphone data selection.” *Mechanical Systems and Signal Processing* 223 (2025): 112735.

Dai, W, et al. “Instructblip: Towards general-purpose vision-language models with instruction tuning.” *Advances in neural information processing systems, 36, 49250-49267.* 36 (2023): 49250-49267.

Deka, Biplab, et al. “Rico: A Mobile App Dataset for Building Data-Driven Design Applications.” *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology (UIST)*. Association for Computing Machinery (ACM), 2017.

Deng, Xiang, et al. “Mind2Web: Towards a Generalist Agent for the Web.” *Advances in Neural Information Processing Systems (NeurIPS)*. 2023.

Desai, A, et al. “Timevae: A variational auto-encoder for multivariate time series generation.” *arXiv preprint* (2021): 2111.08095 .

Di Girolamo, C, et al. “Which patients are not included in the English Cancer Waiting Times monitoring dataset, 2009–2013? Implications for use of the data in research.” *British Journal of Cancer* 118.5 (2018): 733-737.

Dimopoulou, S, et al. “Mobile Anonymization and Pseudonymization of Structured Health Data for Research.” *Proc. Mobile and Secure Services (MOBISECSERV)* (2022): 1-6.

Dosovitskiy, A, et al. “An image is worth 16x16 words: Transformers for image recognition at scale.” *In International Conference on Learning Representations (ICLR)* (2021).

Drouin, Alexandre, et al. “WorkArena: How Capable Are Web Agents at Solving Common Knowledge Work Tasks?” *International Conference on Machine Learning (ICML)*. PMLR (Proceedings of Machine Learning Research), 2024.

Fang, Chunrong, et al. “Enhanced Crowdsourced Test Report Prioritization via Image-and-Text Semantic Understanding and Feature Integration.” *IEEE Transactions on Software Engineering*. IEEE, 2024.

—. “Enhanced Crowdsourced Test Report Prioritization via Image-and-Text Semantic Understanding and Feature Integration (TIURP).” *IEEE Transactions on Software Engineering* (2024).

Feichtenhofer, C, et al. “SlowFast networks for video recognition.” *In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2019).

Feng, J, et al. “Citygpt: Empowering urban spatial cognition of large language models.” *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining* 2 (2025): 591-602.

Friedrich, F, et al. “Fair Diffusion: Instructing Text-to-Image Generation Models on Fairness.” *arXiv preprint* (2023): arXiv:2302.10893. <https://arxiv.org/abs/2302.1089>.

Fujinami, K, et al. “Evaluating behavior recognition pipeline of laying hens using wearable inertial sensors.” *Sensors* 23.11 (2023): 5077.

Gao, Lian, et al. “MobileViews: A Large-Scale Mobile GUI Dataset.” 2024. <https://arxiv.org/abs/2409.14337>.

Gao, Y, et al. “Retrieval-augmented generation for large language models: A survey.” *arXiv preprint* (2023): arXiv:2312.10997.

Gebru, T, et al. “Datasheets for datasets.” *Communications of the ACM* 64.12 (2021): 86–92.

Ghalavand, H, et al. “Common data quality elements for health information systems: A systematic review.” *BMC Medical Informatics and Decision Making* 24.1 (2024): 243.

Ghosh, R, G and P Jiang. “Artificial intelligence framework to identify spatial patterns of cancer immunotherapy outcomes from multiplexed proteomics imaging.” *Journal for Immunotherapy of Cancer* 13.2 (2025): A1243.

Girten, W. *Building Modern Data Applications Using Databricks Lakehouse: Develop, Optimize, and Monitor Data Pipelines in Databricks*. 1. Birmingham, U.K.: Packt Publishing Ltd., 2024.

Godahewa, R, et al. “Monash time series forecasting archive.” *arXiv preprint* (2021): 2105.06643.

Gou, Boyu, et al. “Navigating the Digital World as Humans Do: Universal Visual Grounding for GUI Agents.” 2024. <https://arxiv.org/abs/2410.05243>.

Gururangan, S, et al. “Don't stop pretraining: Adapt language models to domains and tasks. .” *Proceedings of the 58th Annual Meeting of the Association for Computational Ling* (2020): pp. 8342-8360.

Hanocka, R, et al. “Meshcnn: a network with an edge.” *ACM Transactions on Graphics (ToG)* 38.4 (2019): 1-12.

Harada, R. “Simple, yet efficient conformational sampling methods for reproducing/predicting biologically rare events of proteins.” *Bulletin of the Chemical Society of Japan* 91.9 (2018): 1436–1450.

He, Xuehai, et al. “PathVQA: 30000+ Questions for Medical Visual Question Answering.” 2020. <https://arxiv.org/abs/2003.10286>.

Henry, Felix, et al. “OmniGUI: Benchmarking GUI Agents in Omni-Modal Smartphone Environments.” 2026. <https://arxiv.org/abs/2605.18758>.

Ho, X, et al. “Format matters: The robustness of multimodal LLMs in reviewing evidence from tables and charts.” *Proceedings of the AAAI Conference on Artificial Intelligence* 40.37 (2026): 31014-31022.

—. “Table-Text Alignment: Explaining Claim Verification Against Tables in Scientific Papers.” *arXiv preprint* (2025): 2506.10486.

Hongjin, S, U, et al. “Selective annotation makes language models better few-shot learners.” *In The Eleventh International Conference on Learning Representations* (2022).

Hottong, M., M. Sperling and C. Müller. “Using Auto-ML on Synthetic Point Cloud Generation.” *Applied Sciences* 14 (2024): 742.

Hou, X, Z Chen and Y Shang. “A Cross-modal consistency-driven quality-aware dynamic fusion framework for emergency damage assessment.” *IEEE Access* 14 (2026): 70654–70672.

Houck, K, M. “SIERA-ex (Single-Island Endemic Representativeness Analysis for Ex Situ Collections): a Kaua‘i-Based Plant Conservation Model.” *ProQuest Dissertations & Theses Global* (2025).

Hsiao, Yu-Cheng, et al. “ScreenQA: Large-Scale Question-Answer Pairs Over Mobile App Screenshots.” 2022. <https://arxiv.org/abs/2209.08199>.

Hu, Z, M Rostami and J Thomason. “Multi-modal Synthetic Data Training and Model Collapse: Insights from VLMs and Diffusion Models.” *arXiv preprint* (2025): arXiv:2505.08803.

Hutchinson, B, et al. “Towards accountability for machine learning datasets: Practices from software engineering and infrastructure.” *In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)* (2021).

Jacques-Silva, G, et al. “Unified lineage system: Tracking data provenance at scale.” *In Companion of the 2025 International Conference on Management of Data* (2025): 457-470.

Jang, Yuenan, et al. “VideoWebArena: Evaluating Long Context Multimodal Agents with Video Understanding Web Tasks.” 2024. <https://arxiv.org/abs/2410.19100>.

Johnson, Alistair E. W., et al. “MIMIC-CXR, a De-Identified Publicly Available Database of Chest Radiographs with Free-Text Reports.” *Scientific Data* 6 (2019).

Jonnala, S, P, K Parida and N, M Thomas. “EU AI Act underrepresented and insufficient to address the risk and vulnerabilities of generative AI.” *International Journal of Business Analytics* 12.1 (2025): 1-27.

Kantaros, Antreas, et al. “From Mathematical Modeling and Simulation to Digital Twins: Bridging Theory and Digital Realities in Industry and Emerging Technologies.” *Applied Sciences* (2025): vol. 15, p. 9213.

Karampela, M, S Ouhbi and M Isomursu. “Exploring users’ willingness to share their health and personal data under the prism of the new GDPR: implications in healthcare,.” *in Proc. IEEE Eng. Med. Biol. Soc. Annu. Int. Conf.* (2019): 6509–6512.

Karpukhin, V, et al. “Dense passage retrieval for open-domain question answering.” *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNL* November (2020).

Koh, Jing Yu, et al. “VisualWebArena: Evaluating Multimodal Agents on Realistic Visual Web Tasks.” *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 2024.

Kong, Y, et al. “Time-MQA: Time Series Multi-Task Question Answering with Context Enhancement.” *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics* 1 (2025): 29736-29753.

Kwilinski, A and O Reznik. “Governance of artificial intelligence technologies and systems in the EU and Ukraine: Legal foundations and institutional mechanisms.” *Forum Scienciae Oeconomia* 13.3 (2025).

Lacoste, A, et al. “Quantifying the carbon emissions of machine learning.” *arXiv preprint* (2019): arXiv:1910.09700.

Lasim, O, U, E, W Ansah and D Apaak. “Maternal and child health data quality in health care facilities at the Cape Coast Metropolis, Ghana.” *BMC Health Services Research* 22.1 (2022): 1102.

Lau, Jason J., et al. “A Dataset of Clinically Generated Visual Questions and Answers About Radiology Images.” *Scientific Data* 5 (2018).

Lazaridou, A, et al. “Mind the gap: Assessing temporal generalization in neural language models.” *Advances in Neural Information Processing Systems (NeurIPS)* 34 (2021).

Le Roux, R, et al. “Le Roux et al. (2026). A hybrid game engine–generative AI framework for overcoming data scarcity in open-pit crack detection. Machine Learning and Knowledge Extraction, 8(4), 99. https://doi.org/10.3390/make8040099.” *Machine Learning and Knowledge Extraction* 8.4 (2026): 99.

Lee, K, et al. “Deduplicating training data makes language models better.” *In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)* (2022).

Lehman, E, et al. “Do we still need clinical language models?.” *In Conference on health, inference, and learning* PMLR.June (2023): pp. 578-597.

Lewis, P, et al. “Retrieval-augmented generation for knowledge-intensive nlp tasks. .” *Advances in neural information processing systems* 33 (2020): 9459-9474.

Lhoest, Q, et al. “Datasets: A community library for natural language processing: system demonstrations.” *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processi* (2021): pp. 175-184.

Li, Bohao and Rui Wang. “SEED-Bench: Benchmarking Multimodal LLMs with Generative Comprehension.” 2023.

Li, H, et al. “BRIDGE: Bootstrapping Text to Control Time-Series Generation via Multi-Agent Iterative Optimization and Diffusion Modeling.” *arXiv preprint* (2025): 2503.02445.

Li, Kaixin, et al. “ScreenSpot-Pro: GUI Grounding for Professional High-Resolution Computer Use.” arXiv, 2025. <https://arxiv.org/abs/2504.07981>.

Li, M, et al. “M3DBench: Let's Instruct Large Models with Multi-modal 3D Prompts.” *arXiv* (2023): 2312.10763.

Li, T, et al. “Predicting Stress–Strain Curve with Confidence: Balance Between Data Minimization and Uncertainty Quantification by a Dual Bayesian Model.” *Polymers* 17.4 (2025): 550.

Li, Wei, et al. “On the Effects of Data Scale on UI Control Agents.” *Advances in Neural Information Processing Systems (NeurIPS)*. 2024.

Li, Yang, et al. “Mapping Natural Language Instructions to Mobile UI Action Sequences.” *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics (ACL), 2020. 8198–8210.

—. “Widget Captioning: Generating Natural Language Description for Mobile User Interface Elements.” *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics (ACL), 2020. 5495–5510.

Lin, J, et al. “Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations.” *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval* (2021): 2356-2362.

Liu, H, et al. “Time-mmd: Multi-domain multimodal dataset for time series analysis.” *Advances in Neural Information Processing Systems* 37 (2024): 77888-77933.

—. “Visual instruction tuning.” *Advances in neural information processing systems* 36 (2023): 34892-34916.

Liu, J, et al. “What makes good in-context examples for GPT-3?” *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd workshop on knowledge extraction and integration for deep learning architectures* May (2022): 100-114.

Liu, T, F Wu and V Kumar. “A sensor-based IoT data collection and marine economy collaborative innovation method.” *Computational Intelligence and Neuroscience* 3421999 (2022).

Liu, Xiaoran and David Istvan. “AI Simulation by Digital Twins: Systematic Survey, Reference Framework, and Mapping to a Standardized Architecture.” *Software and Systems Modeling* (2025): 1-26.

Liu, Yuan, et al. “MMBench: Is Your Multi-Modal Model an All-Around Player?” *European Conference on Computer Vision (ECCV)*. Springer, 2024. 216–233.

Liu, Z, G, et al. “Conditional-TimeGAN for realistic and high-quality appliance trajectories generation and data augmentation in nonintrusive load monitoring.” *IEEE transactions on instrumentation and measurement* 73 (2024): 1-15.

Liu, Zhen, et al. “Seeing Is Believing: Vision-Driven Non-Crash Functional Bug Detection for Mobile Apps.” *IEEE Transactions on Software Engineering* 51.12 (2025): 3452–3466.

Lottick, K, et al. “Energy usage reports: Environmental awareness as part of algorithmic accountability.” *arXiv preprint* (2019): arXiv:1911.08354.

Lu, Pan, et al. “MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts.” *International Conference on Learning Representations (ICLR)*. 2024.

Lu, Qi, et al. “GUI Odyssey: A Comprehensive Dataset for Cross-App GUI Navigation on Mobile Devices.” 2024. <https://arxiv.org/abs/2406.08451>.

Lu, X, et al. “SCITAB: A challenging benchmark for compositional reasoning and claim verification on scientific tables.” *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (2023): pp. 7787-7813.

Lu, Y, et al. “Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity.” *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* 1.May (2022): 8086-8098.

Mandolini, M., et al. “A cost modelling methodology based on machine learning for engineered-to-order products.” *Engineering Applications of Artificial Intelligence, 136, 108957.* 136 (2024): 108957.

Martinez-Gonzalez, P., et al. “Unrealrox: an extremely photorealistic virtual reality environment for robotics simulations and synthetic data generation. Virtual Re.” *Virtual Reality* 24.2 (2020): 271-288.

Martins, P, H, et al. “Eurollm-9b: Technical report.” *arXiv preprint* (2025): 2506.04079.

Massenon, Romain, et al. “Toward an Automated Cross-Multimodal Verification of Mobile App Bug Fixes Integrating User Feedback, Developer Responses, Changelogs, and UI Visual Analysis.” *Information and Software Technology* 191 (2026).

Mazhar, A, A, et al. “AI-powered detection of cyberbullying in short-form video content: A hybrid deep learning framework.” *PLoS ONE* 21.2 (2026): e0338799.

Mishra, S, et al. “Cross-task generalization via natural language crowdsourcing instructions.” *In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* 1 (2022): 3470-3487.

Miyamoto, M and A Kasuga. “Enterprise Data Science Platform: A Unified Architecture for Federated Data Access.” *in Proc. IEEE Int. Conf. Big Data* (2025): 7395–7404.

Mo, S, et al. “MultiIoT: Benchmarking Machine Learning for the Internet of Things.” *arXiv preprint* (2023): 2311.06217.

Mourby, M, et al. “Are ‘pseudonymised’ data always personal data? Implications of the GDPR for administrative data research in the UK.” *Comput. Law Secur. Rev.* 34.2 (2018): 222–233.

Mousavi, A, et al. “E-CaTCH: Event-centric cross-modal attention with temporal consistency and class-imbalance handling for misinformation detection.” *IEEE Transactions on Computational Social Systems* (2026): 1-14.

Mu, Jiachen, et al. “GUI-360°: A Comprehensive Dataset and Benchmark for Computer-Using Agents.” 2025. <https://arxiv.org/abs/2511.04307>.

Mwubahimana, B, et al. “C2FNet: Cross-Probabilistic Weak Supervision Learning for High-Resolution Land Cover Enhancement.” *IEEE Transactions on Geoscience and Remote Sensing* (2025).

Najjar, Y, S, H and A Abu-Shamleh. “Performance evaluation of a large-scale thermal power plant based on the best industrial practices.” *Scientific Reports* 10 (2020): 20661.

Nashaat, M, et al. “Asterisk: Generating large training datasets with automatic active supervision.” *ACM Transactions on Data Science* 1.2 (2020): 1-25.

Naziroglu, R, E, et al. “Semi-automatic bowel wall thickness measurements on MR enterography in patients with Crohn's disease.” *British Journal of Radiology* 90.1074 (2017): 20160654.

Nguyen, D, C, et al. “Federated learning for industrial Internet of Things in future industries.” *IEEE Wireless Communications* 28.6 (2021): 192–199.

O'Shea, T, J, T Roy and T, C Clancy. "Over-the-air deep learning based radio signal classification." *IEEE Journal of Selected Topics in Signal Processing* 12.1 (2018): 168-179.

Park, J. "Sequential and temporal analysis of human-generated data (Doctoral dissertation)." *ProQuest Dissertations & Theses Global* ISBN 9798607316488 (2019).

Park, J, S, et al. "Generative agents: Interactive simulacra of human behavior." *Proceedings of the 36th annual acm symposium on user interface software and technology* October (2023): 1-22.

Pastrián, D, et al. "Evolutionary Multi-Objective Prompt Learning for Synthetic Text Data Generation with Black-Box Large Language Models." *Applied Sciences* 16.8 (2026): p. 3623.

Patterson, D, et al. "Carbon considerations for large AI models." *arXiv preprint* (2021): arXiv:2104.10350.

Pham, T, D, et al. "SyNC: Balancing Fidelity and Diversity of Synthetic Data Representations in CLIP-based Few-Shot Learning via Neural Collapse." *under review as a conference paper at ICLR* (2026). <https://openreview.net/forum?id=nsUVWQiboS>.

Phinias, R, M Muhanga and J Malago. "Variability in perceived healthcare data quality across Tanzanian regional referral hospitals." *Health Science Reports* 9.3 (2026).

Plattner, N, et al. "An Infinite Swapping Approach to the Rare-Event Sampling Problem." *Journal of Chemical Physics* 135.13 (2011): 134111.

Pu, R, et al. "NOTO: Noise-Tolerate Evidential Learning for Open-Set Cross-modal Retrieval." *IEEE Transactions on Image Processing* (2026).

Pu, Y, et al. "Facial Data Minimization: Shallow Model as Your Privacy Filter." *IEEE Trans. Dependable Secure Comput.* 22.5 (2025): 5186–5202.

Qi, C, R, et al. "Pointnet: Deep learning on point sets for 3d classification and segmentation." *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017): pp. 652-660.

Qi, C. R, et al. "Pointnet++: Deep hierarchical feature learning on point sets in a metric space." *Advances in neural information processing systems* 30 (2017).

Qian, J, et al. "Timeldm: Latent diffusion model for unconditional time series generation. ." *arXiv preprint* (2024): 2407.04211.

Raffel, C, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *Journal of machine learning research* 21.140 (2020): 1-67.

Rajaraman, S and S, K Antani. "Modality-Specific Deep Learning Model Ensembles Toward Improving TB Detection in Chest Radiographs." *IEEE Access* 8 (2020): 27318–27326.

Raji, I, D, et al. "AI and the everything in the whole wide world benchmark." *In Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track* (2021).

Ramos, M, M, et al. "EuroLLM-22B: Technical Report." *arXiv preprint* (2026): 2602.05879.

Raso, E, et al. "Anonymization and pseudonymization of FHIR resources for secondary use of healthcare data." *IEEE Access* 12 (2024): 44929–44939.

Rawles, Christopher, et al. “Android in the Wild: A Large-Scale Dataset for Android Device Control.” *Advances in Neural Information Processing Systems (NeurIPS)*. 2023.

Rebuffi, S, A, et al. “Learning to Watermark in the Latent Space of Generative Models.” *arXiv preprint* (2026): arXiv:2601.16140. <https://arxiv.org/abs>.

Robertson, S and H Zaragoza. “The probabilistic relevance framework: BM25 and beyond.” *Now Publishers Inc* 4 (2009).

Rokem, A, et al. “Open-source models for development of data and metadata standards.” *Patterns* 6.7 (2025): 101316.

Sanh, V, et al. “ Multitask prompted training enables zero-shot task generalization. .” *arXiv preprint* (2021): arXiv:2110.08207.

Sarthi, P, S Abdullah and A Tuli. “RAPTOR: recursive abstractive processing 505 for tree-organized retrieval. .” *arXiv preprint* (2024): arXiv:2401.18059, 506.

Schepaschenko, D, et al. “Development of a global hybrid forest mask through the synergy of remote sensing, crowdsourcing and FAO statistics.” *Remote Sensing of Environment* 162 (2015): 208-220.

Schick, T, et al. “Toolformer: Language models can teach themselves to use tools. .” *Advances in neural information processing systems* 36 (2023): 68539-68551.

Schuhmann, C, et al. “Laion-400m: Open dataset of clip-filtered 400 million image-text pairs.” *arXiv preprint* (2021): 2111.02114.

—. “Laion-5b: An open large-scale dataset for training next generation image-text models.” *Advances in neural information processing systems* 35 (2022): 25278-25294.

Schwartz, R, et al. “Green AI.” *Communications of the ACM* 63.12 (2020): 54–63.

Shi, F, et al. “Large language models can be easily distracted by irrelevant context.” *International Conference on Machine Learning* PMLR.July (2023): pp. 31210-31227.

Si, Chenglei, et al. “Design2Code: How Far Are We from Automating Front-End Engineering?” 2024. <https://arxiv.org/abs/2403.03163>.

Sim, I, et al. “Ontology-based federated data access to human studies information.” *Proc. AMIA Annu. Symp* (2012): 856–865.

Singh, S, et al. “Unlocking Model Insights: A Dataset for Automated Model Card Generation.” *arXiv* (2023).

Sorscher, B, et al. “Beyond neural scaling laws: beating power law scaling via data pruning.” *Advances in Neural Information Processing Systems (NeurIPS)* 35 (2022).

Strubell, E, A Ganesh and A McCallum. “Energy and policy considerations for deep learning in NLP.” *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)* (2019).

Sun, Pei, et al. “Scalability in Perception for Autonomous Driving: Waymo Open Dataset.” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020.

Takasan, K and M Iiyama. “Improving fishing ground estimation with weak supervision and meta-learning.” *PLOS ONE* 20.4 (2025): e0321116.

Tan, Xuchen, et al. “ImageR: Enhancing Bug Report Clarity by Screenshots.” 2025. <https://arxiv.org/abs/2505.01925>.

Tao, R, et al. “Fine-grained cross-modal semantic consistency in natural conservation image data from a multi-task perspective.” *Sensors* 24.10 (2024): 3130.

Tian, C, et al. “Learning a robust RGB-Thermal detector for extreme modality imbalance.” *Pattern Recognition Letters* 196 (2025): 1–8.

Touvron, H, et al. “ Llama: Open and efficient foundation language models.” *arXiv preprint* (2023): arXiv:2302.13971.

Van der Velde, R, et al. “Twelve years of profile soil moisture and temperature measurements in Twente, the Netherlands.” *Earth System Science Data* 15.4 (2023): 1889–1910.

Vangay, P, et al. “Microbiome metadata standards: Report of the National Microbiome Data Collaborative's workshop and follow-on activities.” *mSystems* 6.1 (2021).

Villalobos, P, et al. “Will we run out of data? Limits of LLM scaling based on human-generated data.” *arXiv preprint* (2024): arXiv:2211.04325.

Voronov, A, L Wolf and M Ryabinin. “Mind your format: Towards consistent evaluation of in-context learning improvements.” *Findings of the Association for Computational Linguistics* (2024): 6287-6310.

Wang, Bryan, et al. “Screen2Words: Automatic Mobile UI Summarization with Multimodal Learning.” *Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology (UIST)*. Association for Computing Machinery (ACM), 2021.

Wang, Y, et al. “Internvid: A large-scale video-text dataset for multimodal understanding and generation.” *arXiv preprint* (2023): 2307.06942.

Webber, R, J, et al. “Practical rare event sampling for extreme mesoscale weather.” *Chaos: An Interdisciplinary Journal of Nonlinear Science* 29.5 (2019): 053109.

Wei, J, et al. “Finetuned language models are zero-shot learners.” *arXiv preprint* (2021): arXiv:2109.01652.

Weizman, L, et al. “Plexiform neurofibroma tissue classification.” *Proceedings of SPIE* 7962.1 (2011): 79623K.

Wen, S, et al. “Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation.” *arXiv preprint* (2026): arXiv:2503.14905v2.

Xie, Ming, et al. “NiCro: Purely Vision-Based, Non-Intrusive Cross-Device and Cross-Platform GUI Testing.” 2023. <https://arxiv.org/abs/2305.14611>.

Xie, Tianbao, et al. “OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments.” *Advances in Neural Information Processing Systems (NeurIPS)*. 2024.

Xue, Tian, et al. “An Illusion of Progress? Assessing the Current State of Web Agents.” 2025. <https://arxiv.org/abs/2504.01382>.

Yaghi, A., et al. “3dgenie: synthetic point clouds for semantic segmentation in realistic virtual environments.” *Multimedia Tools and Applications* (2025): 1-33.

Yang, C, et al. “Chartmimic: Evaluating lmm’s cross-modal reasoning capability via chart-to-code generation.” *arXiv preprint* (2025): 2406.09961.

Yang, Jing, et al. “SWE-Bench Multimodal: Do AI Systems Generalize to Visual Software Domains?” 2024. <https://arxiv.org/abs/2410.03859>.

Yang, Y, et al. “Exploiting cross-modal prediction and relation consistency for semisupervised image captioning.” *IEEE Transactions on Cybernetics* 54.2 (2022): 890-902.

Yao, S, et al. “React: Synergizing reasoning and acting in language models.” *arXiv preprint* (2022): arXiv:2210.03629.

Yoon, Jinsung, Daniel Jarrett and Mihaela Van der Schaar. “Time-series generative adversarial networks.” *Advances in neural information processing systems* 32 (2019).

Yu, Fisher, et al. “BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning.” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020. 2636–2645.

Yu, Shengcheng, et al. “Semi-Supervised Crowdsourced Test Report Clustering via Screenshot–Text Binding Rules.” *Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering (FSE)*. ACM, 2024.

Yu, Shuai, et al. “Mobile App Crowdsourced Test Report Consistency Detection via Deep Image-and-Text Fusion Understanding.” *IEEE Transactions on Software Engineering* 49.8 (2023): 4115–4134.

—. “Prioritize Crowdsourced Test Reports via Deep Screenshot Understanding.” *Proceedings of the 43rd IEEE/ACM International Conference on Software Engineering (ICSE)*. IEEE Computer Society / ACM, 2021. 946–956.

Yuan, Xinyu and Yan Qiao. “Diffusion-ts: Interpretable diffusion for general time series generation.” *arXiv preprint* (2024): 2403.01742.

Yuan, Y, et al. “SynEval: A Multidimensional Framework for Evaluating Synthetic Data on Fidelity, Utility, Diversity, and Privacy.” IEEE Symposium on Security and Privacy, 2026. <https://www.ieee-security.org/TC/SP2026/downloads/posters/sp2026posters-final90.pdf>.

Yue, Xiang, et al. "MMMU: A Massive Multi-Discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024. 9556–9567.

Yun, Seonghyeon, et al. “Web2Code: A Large-Scale Webpage-to-Code Dataset and Evaluation Framework for Multimodal LLMs.” *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*. 2024.

Zellers, R, et al. “Merlot: Multimodal neural script knowledge models.” *Advances in neural information processing systems* 34 (2021): 23634-23651.

Zhang, D, et al. “Multimodal distillation and fusion for enhanced age-related macular degeneration classification.” *IEEE Journal of Biomedical and Health Informatics* (2026).

Zhang, H, B, et al. “A survey of deep learning-based action quality assessment: methods and applications.” *Journal of Big Data,* 13.1 (2026): 70.

Zhang, W, et al. “Applications of the Cross-Entropy Method to Importance Sampling and Optimal Control of Diffusions.” *SIAM Journal on Scientific Computing* 36.6 (2014): A2654–A2672.

Zhang, Xinyao, et al. “Proactive Detection of GUI Defects in Multi-Window Scenarios via Multimodal Reasoning.” 2026. <https://arxiv.org/abs/2604.19081>.

Zhang, Y and Y, F Zhao. “A Web-based Automated Manufacturability Analyzer and Recommender for Additive Manufacturing (MAR-AM) via a Hybrid Machine Learning Model.” *Expert Systems with Applications* (2022).

Zhao, B, K, R Mopuri and H Bilen. “Dataset condensation with differentiable siamese augmentation.” *In Proceedings of the 38th International Conference on Machine Learning (ICML)* (2021).

Zhao, W, et al. “Advancing performance via a systematic application of research and industrial best practice.” *Proceedings of the ACM on Programming Languages* 9.OOPSLA2 (2025): 4008–4034.

Zheng, Y, X Xie and W, Y Ma. “GeoLife: A collaborative social networking service among user, location and trajectory.” *IEEE Data Eng. Bull* 33.2 (2010): 32-39.

Zhong, H, et al. “HONEYBEE: Efficient Role-based Access Control for Vector Databases via Dynamic Partitioning.” *Proceedings of the ACM on Management of Data* 4.1 (2026): 11.

Zhou, Shuyan, et al. “WebArena: A Realistic Web Environment for Building Autonomous Agents.” *International Conference on Learning Representations (ICLR)*. International Conference on Learning Representations, 2024.

Zhu, B, Z Hu and Z Lu. “LiDAR-Free 3D Auto-Labeling via Radar–Visual Spatio-Temporal Consistency.” *Sensors* 26.10 (2026): 2956.

Zhu, Z and I Brilakis. “Comparison of optical sensor-based spatial data collection techniques for civil infrastructure modeling.” *Journal of Computing in Civil Engineering* 23.3 (2009): 170–177.

Zolanvari, M. “Addressing pragmatic challenges in utilizing AI for security of industrial IoT (Doctoral dissertation).” *ProQuest Dissertations & Theses Global* (2021).

