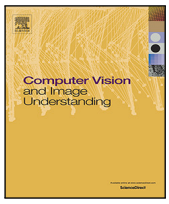
Computer Vision and Image Understanding 249 (2024) 104205

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu



M^3A: A multimodal misinformation dataset for media authenticity analysis

Qingzheng Xu [a], Huiqiang Chen [b], Heming Du [a], Hu Zhang [a], Szymon Łukasik [c], Tianqing Zhu [d], Xin Yu [a,*]

[a] The University of Queensland, 280-284 Sir Fred Schonell Dr, St Lucia QLD 4067, Australia
[b] The University of Technology Sydney, 15 Broadway, Ultimo NSW 2007, Australia
[c] AGH University of Science and Technology, al. Adama Mickiewicza 30, 30-059 Kraków, Poland
[d] City University of Macau, Avenida Padre Tomás Pereira Taipa, Macau, China

ARTICLE INFO

MSC:
91D30
68U10
68T50
68T45

Keywords:
Misinformation detection
Media authenticity
Multimodal dataset

ABSTRACT

With the development of various generative models, misinformation in news media becomes more deceptive and easier to create, posing a significant problem. However, existing datasets for misinformation study often have limited modalities, constrained sources, and a narrow range of topics. These limitations make it difficult to train models that can effectively combat real-world misinformation. To address this, we propose a comprehensive, large-scale Multimodal Misinformation dataset for Media Authenticity Analysis (M^3A), featuring broad sources and fine-grained annotations for topics and sentiments. To curate M^3A, we collect genuine news content from 60 renowned news outlets worldwide and generate fake samples using multiple techniques. These include altering named entities in texts, swapping modalities between samples, creating new modalities, and misrepresenting movie content as news. M^3A contains 708K genuine news samples and over 6M fake news samples, spanning text, images, audio, and video. M^3A provides detailed multi-class labels, crucial for various misinformation detection tasks, including out-of-context detection and deepfake detection. For each task, we offer extensive benchmarks using state-of-the-art models, aiming to enhance the development of robust misinformation detection systems.

1. Introduction

In light of recent advancements in generative technologies (Brown et al., 2020; Wang et al., 2021, 2022; Sauer et al., 2023; Ghosal et al., 2023; Khachatryan et al., 2023; Wu et al., 2024) and foundation models (Radford et al., 2021; Li et al., 2022; Girdhar et al., 2023; Zhu et al., 2024), one can easily generate high-fidelity fake news that misleads the public and causes significant societal trust issues (Abdelnabi et al., 2022a).

Existing misinformation detection methods are generally trained based on artificially synthesized misinformation datasets, such as multimodal misinformation datasets (Jaiswal et al., 2017; Sabir et al., 2018; Tan et al., 2020; Müller-Budack et al., 2020; Shivangi Aneja and Nießner, 2023; Luo et al., 2021; Rui Shao and Liu, 2023) and deepfake datasets (Yang et al., 2019; Dolhansky et al., 2019; Rossler et al., 2019; Li et al., 2020; Jiang et al., 2020; Huang et al., 2021). However, all these datasets are limited to at most two modalities, lacking variety in data sources, and topics. As a result, detection models trained on them are insufficiently prepared to tackle the complexities of real-world misinformation challenges. There is an urgent need for a more comprehensive and diverse misinformation dataset.

This paper proposes M^3A, a comprehensive large-scale multimodal misinformation dataset. M^3A enables a variety of misinformation detection tasks, as detailed in Fig. 1. This dataset includes authentic news content from 60 prominent news outlets worldwide, featuring texts, images, audio, and videos. The outlets are carefully selected for their reputation and trustworthiness, ensuring they are widely recognized as credible sources of information. M^3A covers a broad range of topics such as politics, technology, and entertainment, reflecting the diversity of contemporary news. By incorporating varying text lengths, image resolutions, speech lengths, and video durations, M^3A enhances its complexity, making it an invaluable resource for training advanced misinformation detection models. This extensive dataset equips researchers with the necessary data to develop models capable of tackling the multifaceted challenges of misinformation in today's media landscape.

We create fabricated news content through multiple strategies, *i.e.*, (1) ***Named Entity Manipulation (NEM)***, which involves replacing named entities such as person, location, and organization names in the text; (2) ***Multimodality Mismatching (MM)***, where we replace the text, image, audio, or video of a sample with another from a

* Corresponding author.
E-mail address: xin.yu@uq.edu.au (X. Yu).

https://doi.org/10.1016/j.cviu.2024.104205
Received 31 May 2024; Received in revised form 6 October 2024; Accepted 8 October 2024
Available online 15 October 2024

Table 1
Comparison of existing multimodality misinformation datasets. In this table, "OOC" refers to out-of-context issue detection, "DF det." refers to deepfake detection, and "OOD" refers to out-of-distribution testing.

Dataset	Source (# News outlets)	Size	Data modality				Associated tasks			
			Text	Image	Audio	Video	OOC	DF det.	Fact check	OOD
MAIM (Jaiswal et al., 2017)	Flickr	239k	✓	✓	✗	✗	✓	✗	✓	✗
MEIR (Sabir et al., 2018)	Flickr	57k	✓	✓	✗	✗	✓	✗	✓	✗
NeuralNews (Tan et al., 2020)	GoodNews (1)	128k	✓	✓	✗	✗	✓	✗	✓	✗
TamperedNews (Müller-Budack et al., 2020)	BreakingNews (4)	776k	✓	✓	✗	✗	✓	✗	✓	✗
COSMOS (Shivangi Aneja and Nießner, 2023)	18 News outlets	453k	✓	✓	✗	✗	✓	✗	✗	✗
NewsCLIPpings (Luo et al., 2021)	VisualNews (4)	988k	✓	✓	✗	✗	✓	✗	✓	✓
DGM^4 (Rui Shao and Liu, 2023)	VisualNews (4)	239k	✓	✓	✗	✗	✓	✓	✓	✗
M^3A(Ours)	60 News outlets	7m	✓	✓	✓	✓	✓	✓	✓	✓

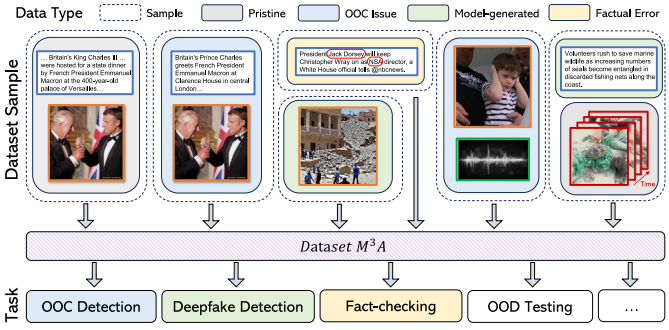


Fig. 1. Demonstration of M^3A Application. M^3A contains data generated in various methods, facilitating multiple misinformation detection tasks such as OOC detection, deepfake detection, fact-checking, OOD testing and so on.

different news sample based on modality similarity; (3) ***Text-driven Multimodality Generation (TMG)***, using pre-trained large language models to generate images, videos, and audio based on the text; (4) ***Multimodality-driven Text Generation (MTG)***, using GPT-like APIs to generate text based on images, audio, and videos; and (5) ***Movie to News (M2N)***, pairing movie stills, sounds, or clips with real news content and model-generated texts to create misinformation. As illustrated in Fig. 2, these methods enhance the diversity of M^3A.

Each sample in our dataset is enriched with detailed annotations, enabling us to set benchmarks for different misinformation detection tasks, as outlined in Fig. 1. These tasks include (1) Out-of-context (OOC) detection, checking whether any two modalities of the input are consistent; (2) Deepfake identification, determining if any input modality is model-generated; (3) Fact-checking, verifying the accuracy of each modality against the true information database of M^3A; and (4) Out-of-distribution (OOD) testing, assessing model robustness on different data. We deploy state-of-the-art models for these benchmarks and report their performance using corresponding evaluation metrics. The experimental results reveal the complexity and diversity of M^3A, highlighting its inherent challenges. The major contributions of this research are summarized as follows:

- We present M^3A, the first comprehensive large-scale multimodal misinformation dataset with news samples in text, image, audio, and video formats from reputable news outlets, addressing limitations in misinformation generation, data modality, scale, and topic diversity.
- M^3A includes multi-class annotations essential for various misinformation detection tasks, such as out-of-context detection, deepfake identification, and fact-checking.
- We propose benchmarks for M^3A tailored to various misinformation detection tasks, utilizing state-of-the-art models and out-of-distribution testing to support further research in misinformation detection.

2. Related works

2.1. Multimodal misinformation dataset

As shown in Table 1, current multimodal misinformation datasets are primarily limited to text and image modalities. For instance, Jaiswal et al. (2017) creates misinformation through simple caption swaps. Sabir et al. (2018) increases complexity by randomly altering named entities and adding GPS details. Tan et al. (2020) focuses on news articles, combining images with articles and captions but replacing real elements with fabricated ones. Müller-Budack et al. (2020) manipulates named entities based on the comprehensive content of the articles. Luo et al. (2021) leverages advanced language and vision models to automatically generate mismatched image-caption pairs. FACTIFY (Mishra et al., 2022) provides extra supporting documents and images, focusing specifically on fine-grained fact-checking. Shivangi Aneja and Nießner (2023) collects numerous image-caption pairs without annotations and uses a self-supervised technique to match different captions describing the same images. DGM^4 (Rui Shao and Liu, 2023) introduces modifications to both text and images. Papadopoulos et al. (2023) conducts a comparative study on challenges such as out-of-context image-caption pairs, cross-modal named entity inconsistency, and their hybrids, highlighting the issues posed by unimodal biases; while VERITE (Papadopoulos et al., 2024) addresses the unimodal biases by sourcing from reputable fact-checking platforms. Additional datasets (Boididou et al., 2015, 2016; Jin et al., 2017; Wang et al., 2018; Nakamura et al., 2019; Khattar et al., 2019; Shu et al., 2020; Biamby et al., 2022; Nielsen and McConville, 2022; Dufour et al., 2024) have expanded their scope to include real-world content from social media, moving beyond traditional news outlets.

All the aforementioned multimodal datasets contain only text and image modalities. With the rise of short video platforms, news in short video format is becoming increasingly important. However, these datasets lack audio and video modality misinformation. The data sources for these datasets are also quite limited. For example, VisualNews (Liu et al., 2021) includes only four news outlets from the UK and USA (The Guardian, BBC, USA TODAY, and The Washington Post), neglecting news from other regions of the world. Additionally, these datasets contain only general news and lack more granular annotations such as news topics. Moreover, datasets such as TwitterCOMMs (Biamby et al., 2022) and Mumin (Nielsen and McConville, 2022) focus on broader social media content, such as tweets. As this content is user-generated, it requires careful verification before it can be utilized. Compared to formal news sources, social media posts generally have lower credibility, leading to misinformation generated from them being less influential and deceptive.

M^3A overcomes these limitations by collecting news from news outlets worldwide, including Europe, America, Asia, and so on. It features text, image, audio, and video modalities, covering a diverse range of news topics. Furthermore, it provides comprehensive annotations, offering a more realistic and varied dataset for misinformation detection research.

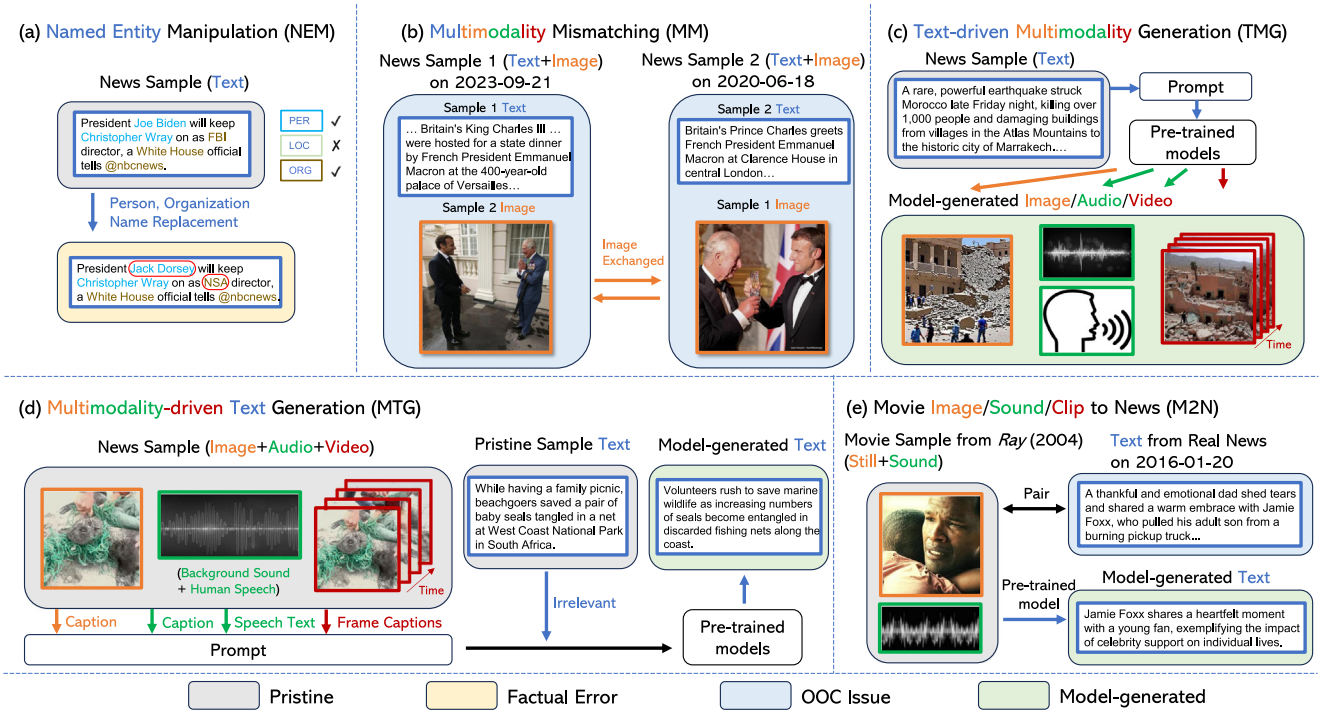


Fig. 2. Illustration of misinformation generation methods in M^3A. (a) Named Entity Manipulation (NEM), (b) Multimodality Mismatching (MM), (c) Text-driven Multimodality Generation (TMG), (d) Multimodality-driven Text Generation (MTG), and (e) Movie to News (M2N), together with four types of data (Pristine, Factual Error, OOC Issue, and Modal-generated).

2.2. Deepfake dataset

Existing deepfake datasets primarily focus on a single modality, such as image or video. For example, UADFV (Yang et al., 2019) consists of 49 real videos paired with 49 deepfake videos, created to highlight inconsistencies in head poses for detecting deepfakes. FaceForensics++ (Rossler et al., 2019) expands on these efforts by offering a large dataset of over 1000 real videos and their manipulated counterparts, created using various face manipulation techniques such as FaceSwap (Chen et al., 2020; Gao et al., 2021), Face2Face (Thies et al., 2016), and NeuralTextures (Thies et al., 2019).

The DeepFake Detection Challenge (Dolhansky et al., 2019) advances deepfake detection technologies by providing over 100,000 labelled videos through a competitive format, utilizing multiple synthesis algorithms to create diverse deepfake videos. CelebDF (Li et al., 2020) includes 590 real videos of celebrities sourced from YouTube, alongside 5639 deepfake videos generated using an improved deepfake synthesis algorithm that addresses issues like low resolution and colour mismatch. DeeperForensics-1.0 (Jiang et al., 2020) includes 60,000 videos of 100 actors, each manipulated to various extents using generative adversarial networks (GAN) (Goodfellow et al., 2014) to simulate real-world scenarios. Face Forensics in the Wild (Zhou et al., 2021) accounts for domain-adversarial factors during the generation of samples. DeepFake MNIST+ (Huang et al., 2021) takes a different approach by using videos from the VoxCeleb dataset (Nagrani et al., 2020), transforming them with deepfake techniques to explore detection in a controlled environment focused on facial animations.

All the above datasets mainly feature talking heads of celebrities or actors instead of realistic news formats, and they lack diversity in both scenarios and content. These datasets often use controlled environments and do not encompass the complexity of real-world news, which involves multiple modalities and more dynamic, unscripted interactions.

M^3A addresses these limitations by using authentic news materials that include talking heads, group scenes, and videos without people. This dataset includes dynamic live reports, interviews, diverse environments, and different reporting styles, such as breaking news and in-depth reports, enhancing the complexity and applicability of misinformation detection research.

3. The M^3A dataset

Current multimodal misinformation datasets typically focus on text and image modalities, limited in their diversity of news sources and topics. To overcome these limitations, we present the M^3A dataset, consisting of a substantial collection of annotated news content spanning text–image pairs and text–image–audio–video pairs.

3.1. Data source

Unlike existing datasets that often have a limited range of data sources, we have significantly expanded the diversity of our dataset by sourcing from 60 prominent media outlets. Inspired by the previous work (Xu et al., 2024), we carefully select these outlets for their reputation and trustworthiness, ensuring they are widely recognized as reliable sources of information. This diverse selection includes international news-focused outlets like ABC News and BBC News, as well as regional sources such as Al Jazeera in the Middle East, The Straits Times and The Times of India in South Asia. Additionally, we include outlets specializing in economic news (e.g., Bloomberg Business), political news (e.g., Politico), entertainment news (e.g., The Sun), and so on. To compile this extensive dataset, we use Instaloader[1] to scrape news posts from these 60 media accounts. Low-quality samples are filtered by removing overly short samples, videos without audio, samples with low inter-modality similarity, and manually removing of low-quality content. In the end, we gather 708,425 original news samples, comprising 526,223 text–image pairs and 182,202 text–image–audio–video pairs. Each news sample is uniquely identified by its ID, formatted as "Publication News Outlet + Publication Time".

[1] https://github.com/instaloader/instaloader

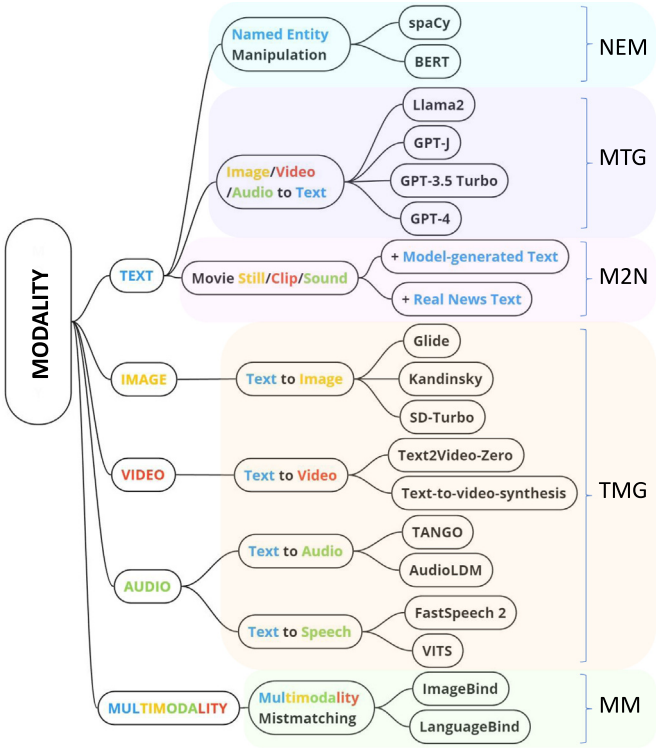


Fig. 3. Methods and models used in M^3A. We utilize various methods and corresponding models to generate fabricated news samples in different modalities. These methods can be summarized as the five types of generation techniques, Named Entity Manipulation (NEM), Multimodality Mismatching (MM), Text-driven Multimodality Generation (TMG), Multimodality-driven Text Generation (MTG), and Movie to News (M2N), as mentioned in Fig. 2.

3.2. Data pre-processing

Although platforms like Instagram are lightweight and fragmented, the length of each text varies according to the writing style of each news outlet. Media outlets such as the Daily Mail and ABC News tend to prefer shorter texts, often with fewer than 20 words. Conversely, outlets like AP News and PBS NewsHour generally share longer texts, many exceeding 300 words. In our experiments, we utilize large language models like CLIP (Radford et al., 2021) and BLIP (Li et al., 2022), which have limitations on input length. For example, CLIP can handle texts up to 77 tokens, which is insufficient for our longest texts. BLIP can manage longer inputs up to 512 tokens, but it is primarily pre-trained on short sentences with fewer than 40 tokens, so directly inputting long sentences may not yield optimal results. By examining the collected data, we observe that not every sentence in a text has a strong connection with the corresponding image. For instance, in texts from The Washington Post, the last sentence often does not closely relate to the news content and typically reads "click this link for more/full story/updates". Therefore, we use BART (Lewis et al., 2020) to overcome the input length limitations. BART summarizes texts by selecting key sentences based on relevance and context, combining them with logical coherence without significantly altering the words, ensuring complete and concise summaries.

We then extract audio from video files using MoviePy.[2] For text–image samples, we employ BLIP-2 (Li et al., 2023) to obtain text and image embeddings. For text–image–audio–video samples, we use ImageBind (Girdhar et al., 2023) and Languagebind (Zhu et al., 2024) to obtain corresponding embeddings of text, image, audio, and video.

3.3. Misinformation generation

As shown in Figs. 2 and 3, we employ five approaches to generate misinformation: Named Entity Manipulation (NEM), Multimodality Mismatching (MM), Text-driven Multimodality Generation (TMG), Multimodality-driven Text Generation (MTG), and Movie to News (M2N). All samples have been meticulously annotated.

Named Entity Manipulation (NEM). This misinformation generation method involves replacing named entities (persons, locations, organizations) within texts as depicted in Fig. 2(a).

To ensure the named entity alteration has a substantial impact on the text meaning, we utilize two named entity recognition models, BERT (Devlin et al., 2019) and spaCy (Honnibal et al., 2020) to build a pool of named entities. Only the intersection of their results for each text will be incorporated into the pool. From the pool, we specifically choose those that occur more than 42 times as viable candidates for replacement, forming the final significant named entities pool.

For each text summary, if it contains a named entity, we replace it with a different one from the significant named entities pool. For the original a text–image sample (T_0, I_0) or text–image–audio–video sample (T_0, I_0, A_0, V_0), if T_0 includes a location name l_0, we randomly select a distinct location name l_1 from the significant named entities pool. Then replace all instances of l_0 throughout the entire text with l_1, resulting in a synthetic text T_1. T_1 is then paired with the original image I_0, or potentially with audio A_0 and video V_0, to produce "location" manipulated sample (T_1, I_0) or (T_1, I_0, A_0, V_0).

Correspondingly, we generate "person" and "organization" manipulated samples. We also create "complete" manipulated samples, in which all named entities in T_0 are modified.

Multimodality Mismatching (MM). This method generates misinformation by re-paring modalities between samples as presented in Fig. 2(b).

For a text–image–audio–video sample (T_0, I_0, A_0, V_0), we apply modality mismatching based on text–text, image–image, audio–audio, and video–video similarity. For instance, to assess image–image similarity, we calculate the cosine similarity between the embedding of image I_0 and all other image embeddings, selecting the image I_1 with the highest similarity score. We ensure a minimum time gap of 30 days between I_0 and I_1 to avoid selecting reports from different news posts covering the same event, ensuring that our image swap can effectively create misinformation. We then replace I_0 with I_1 to create a falsified imaged-changed sample (T_0, I_1, A_0, V_0).

Similarly, we generate text-changed (T_1, I_0, A_0, V_0), audio-changed (T_0, I_0, A_1, V_0), and video-changed (T_0, I_0, A_0, V_1). For a text–image sample (T_0, I_0), in a similar way, we create text-changed (T_1, I_0) and imaged-changed sample (T_0, I_1).

Text-driven Multimodality Generation (TMG). Our data source has a broader range and is not limited to portraits, creating manipulated images by face-swapping is not suitable for M^3A. Thanks to the great advance of generative models, we can adopt pre-trained models to generate falsified images/audio/videos based on texts as detailed in Fig. 2(c).

It is important to note that some of the current generative models, like DALL-E 3,[3] are restricted by content policies, which prevent generating images based on prompts involving celebrities. Therefore, when constructing M^3A, we choose models without such restrictions to generate falsified content.

For image modality, we first explore various text-to-image models, including Glide (Nichol et al., 2021), Kandinsky (Razzhigaev et al., 2023), and SD-Turbo (Sauer et al., 2023). For text T_0, several images are created. The most convincing fabricated image I_1 with the highest

[2] https://github.com/Zulko/moviepy

[3] https://openai.com/index/dall-e-3/

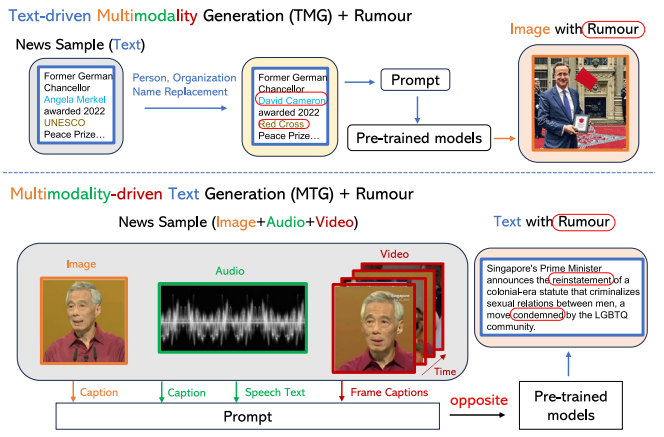


Fig. 4. Illustration of method generation plus rumour. For methods TMG and MTG, in addition to the simple model-generated samples, we also create samples with rumour by swapping named entities in the prompt or altering the prompt to have opposite meanings.

BLIP cosine similarity to T_0 is then selected to create falsified sample (T_0, I_1) or (T_0, I_1, A_0, V_0).

For audio generation, we utilize text-to-audio models, TANGO (Ghosal et al., 2023) and AudioLDM (Liu et al., 2023) as well as text-to-speech models, FastSpeech 2 (Chien et al., 2021) and VITS (Kim et al., 2021a). For video generation, we employ models Text2Video-Zero (Khachatryan et al., 2023) and Text-to-video-synthesis (Wang et al., 2023). Falsified samples (T_0, I_0, A_1, V_0) and (T_0, I_0, A_0, V_1) are then generated.

Multimodality-driven Text Generation (MTG). For this category, we generate fake texts based on images, potentially with audio and videos as illustrated in Fig. 2(d). For an original sample (T_0, I_0) or (T_0, I_0, A_0, V_0), we produce fabricated samples using GPT-like APIs, including Llama 2 (Touvron et al., 2023), GPT-J (Wang and Komatsuzaki, 2021), GPT-3.5 Turbo and GPT-4 (Brown et al., 2020).

To create a prompt, we extract the image caption from I_0 using BLIP (Li et al., 2022). For (T_0, I_0, A_0, V_0), we also extract audio captions using CLAP (Elizalde et al., 2023a,b) and audio speech text using Whisper (Radford et al., 2023) from audio files. By dividing each video into four equal segments, we extracted captions from the first frame of each segment using BLIP and combined them to form the video frame captions. We also record the token number, publication news outlier and publication time of T_0.

Based on image caption or (image caption, audio caption, audio speech text, and video frame captions), use GPT-like APIs to generate falsified texts irrelevant to I_0 with the same token number, publication news outlier and publication time. Similar to method TMG, only the most convincing fabricated text T_1 with the highest BLIP cosine similarity to I_0 or (I_0, A_0, V_0) is selected. T_0 in the original sample is then replaced by T_1 to create a misinformed sample (T_1, I_0) or (T_1, I_0, A_0, V_0).

Generation plus Rumour. In methods TMG and MTG, the generation of new modalities is still based on the existing content of a news sample without adding false information. To make our database more diverse and challenging as well as better simulate the real-world distributions, we deliberately adjust the prompts to generate samples containing rumours.

As shown in Fig. 4, for the TMG plus rumour method, instead of using the original text T_0 as the prompt, we use the corresponding text T_1 from the NEM method (randomly selected from each type). This generates falsified modalities I_1, A_1, and V_1, which are then combined with T_0 to create misinformed samples. For the MTG plus rumour method, instead of prompting GPT-like APIs to generate irrelevant text, we ask them to produce text with opposite meanings while keeping the original named entities intact.

Table 2
Data distribution in M^3A. (T, I) stands for text–image pairs and (T, I, A, V) stands for text–image–audio–video pairs.

Method	Type	(T, I)	(T, I, A, V)	Total
Pristine		526,223	182,202	708,425
NEM	Person	235,657	71,604	307,261
	Location	239,545	81,027	320,572
	Organization	164,195	45,664	209,859
	Complete	423,475	137,561	561,036
MM	Text-changed	526,223	182,202	708,425
	Image-changed	526,223	182,202	708,425
	Audio-changed	–	182,202	182,202
	Video-changed	–	182,202	182,202
TMG	Model-generated	526,223	182,202	708,425
	with rumour	526,223	182,202	708,425
MTG	Model-generated	526,223	182,202	708,425
	with rumour	526,223	182,202	708,425
M2N	Pair with real news	191,116	85,236	199,116
	Model-generated text	191,116	85,236	199,116
Total		5,128,665	2,146,146	7,274,811

Movie to News (M2N). This method fabricates misinformation by presenting movie content as news, as shown in Fig. 2(e).

We observe that many movies contain high-precision scene depictions, such as portrayals of floods and storms in disaster films. These cinematic works often undergo multiple rounds of meticulous review to ensure authenticity, making them ideal candidate sources of fake content. Based on this idea, we collect images from the dataset Movie Stills 2000–2020 Images[4] dataset and YouTube[5] videos made from movie clips. We apply the same process described in methods MM and MTG. We consider these movie stills, sounds, and clips as I_1, A_1, and V_1, respectively. These are then combined with a real news text T_1, selected for its high similarity from M^3A, or a text T_1 generated using GPT-like APIs. In this way, we create new misinformed samples (I_1, T_1) and (I_1, T_1, A_1, V_1).

3.4. Dataset statistics

As illustrated in Table 2, M^3A contains 5,128,665 text–image pairs, including 526,223 pristine samples and 4,602,442 falsified samples. M^3A also includes 2,146,146 text–image–audio–video pairs, including 182,202 pristine samples and 1,963,944 falsified samples.

Fig. 5 provides an extensive overview of the data statistics for M^3A, covering (a) the distribution of data based on different misinformation generation techniques, (b) the allocation of different modalities, (c) the categorization of data according to distinct misinformation detection tasks, and (d) the geographical distribution of the news sources utilized. The figure highlights the dataset's diversity, including a range of generation techniques. Additionally, it shows the distribution of different media types (text, image, audio, video) in M^3A, categorizes the data based on misinformation detection tasks, including out-of-context detection corresponding to methods MM and M2N, deepfake detection linked to methods TMG, MTG, and M2N, and fact-checking associated with method NEM. The figure also highlights the geographical representation of news outlets, showing that USA news outlets hold a dominant position of over 50% in our dataset.

We also conduct a detailed analysis of the news texts, focusing on three aspects: (a) the token count distribution in both the original texts and their summaries generated by BART (Lewis et al., 2020); (b) sentiment analysis using the Twitter-roBERTa-base model (Loureiro et al.,

[4] https://www.kaggle.com/datasets/thevox/movie-stills-20002020-images
[5] https://www.youtube.com/

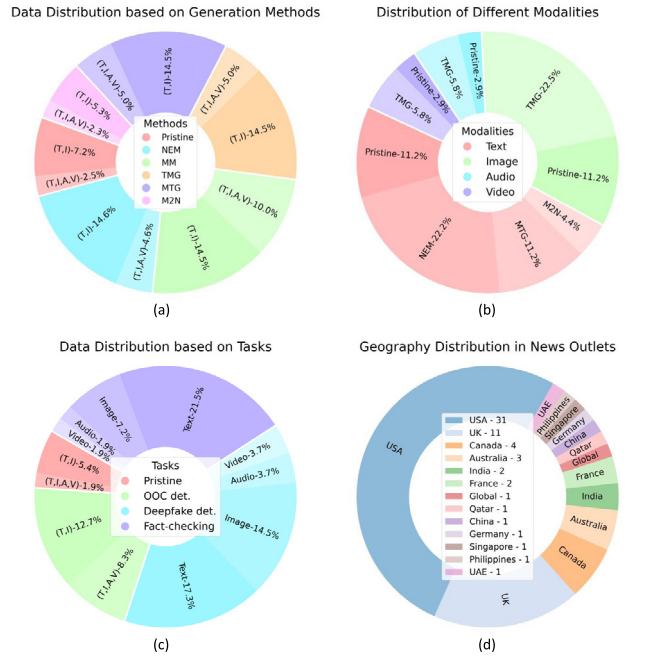


Fig. 5. Data statistics of M^3A. In this figure, "(T, I)" stands for text–image pairs, "(T, I, A, V)" stands for text–image–audio–video pairs, and "det". stands for detection. (a) Data distribution based on different misinformation generation methods; (b) Distribution of different modalities; (c) Data distribution based on various misinformation detection tasks; (d) Geography distribution of selected news outlets.

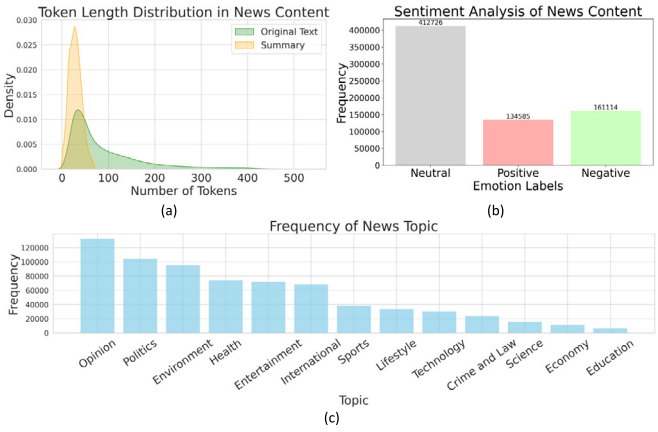


Fig. 6. Data statistics of news content in M^3A. (a) Token length distribution of news contents and their corresponding summaries; (b) Sentiment analysis of news contents; (c) Frequency of news content topics.

2022); and (c) topic distribution using the bart-large model (Lewis et al., 2020). Fig. 6 reveal that: (1) the original texts have a wide range of token counts, while the BART-generated summaries typically contain fewer than 77 tokens, indicating minimal impact on CLIP performance in experiments; (2) the majority of news articles are neutral in tone, with negative articles outnumbering positive ones; and (3) opinion, politics, and environmental topics are the most frequently covered among the 13 categories analysed.

As shown in Fig. 5, M^3A includes a broad array of misinformation. However, it is important to note that M^3A does not encompass every possible form of misinformation. Specifically, we have intentionally excluded content that might be offensive, such as discriminatory, pornographic, or violent material. This decision is made because our data sources are reputable news outlets, which do not typically publish such content. As a result, while M^3A encompasses a broad spectrum of misinformation tactics, it remains within the limits of content deemed acceptable in public discourse. This restriction is crucial for maintaining the dataset's ethical standards and relevance to its intended use.

Additionally, swapping modalities between reports on the same event across different media can generate additional positive samples. Some existing detection tools (Abdelnabi et al., 2022b) based on Google Vision AI often return different URLs, which makes it challenging to identify these positive samples. As part of our future work, we plan to implement an automated process for reviewing and creating such samples to make M^3A more challenging.

4. Experiments

The proposed rich and diverse dataset M^3A enables training various models for different tasks. We propose several potential tasks, test various state-of-the-art baseline models, and document the corresponding benchmark results.

For falsified text–image sample detection, we use BLIP (Li et al., 2022), CLIP (Radford et al., 2021), ViLT (Kim et al., 2021b), and VisualBERT (Li et al., 2019) for classification. We evaluate their performances based on classification accuracy.

- **BLIP** focuses on unified vision–language understanding and generation, employing a multi-task learning approach for image–text matching and language modelling (Li et al., 2022).
- **CLIP** maps images and text into a shared embedding space and aligns visual and textual representations. This model excels in zero-shot learning and has robust generalization capabilities (Radford et al., 2021).
- **ViLT** integrates visual and textual information through a lightweight transformer architecture. It is designed for tasks that require fine-grained alignment between image and text (Kim et al., 2021b).
- **VisualBERT** extends BERT to handle visual inputs by integrating visual and textual data within a unified transformer framework. It is suitable for vision–language tasks like VQA and image captioning (Li et al., 2019).

For falsified text–image–audio–video sample detection, we use ImageBind (Girdhar et al., 2023) and LanguageBind (Zhu et al., 2024). The metrics used are accuracy (Acc), area under the curve (AUC), and average precision (AP).

- **ImageBind** unifies different modalities into a single representation, enabling the model to handle multimodal inputs effectively. It aligns audio, visual, and textual data to create a comprehensive multimodal understanding (Girdhar et al., 2023).
- **LanguageBind** focuses on integrating language with visual and auditory inputs. It enhances the multimodal representation by binding textual descriptions with corresponding audio and visual data, improving the model's ability to process and understand complex multimodal information (Zhu et al., 2024).

4.1. Falsified text-image pairs detection

Zero-shot Misinformation Detection. We first evaluate the performance of BLIP and CLIP models in detecting fabricated news content without fine-tuning on M^3A. Specifically, we compare the similarity scores for each pair, and if the real pair yields a higher score, it is considered a correct prediction. This approach tests the models' generalization capabilities on M^3A.

From the results listed in Table 3(a), CLIP outperforms BLIP in most scenarios, indicating a more robust visual–textual alignment. For instance, CLIP achieves its highest accuracy of 0.697 for detecting misinformation type movie+MM.

Table 3
Performance of misinformed text–image sample detection for different models. (a) Zero-shot prediction performance. (b) Performance after training. The metric used in this table is accuracy. MM results are based on features extracted using BLIP-2.

(a) Zero-shot prediction				(b) Multimodal misinformation detection					
Method	Type	BLIP	CLIP	Method	Type	BLIP	CLIP	ViLT	VisualBERT
NEM	Person	0.488	0.637	NEM	Person	$0.625_{\pm 0.010}$	$0.657_{\pm 0.009}$	$0.509_{\pm 0.009}$	$0.533_{\pm 0.023}$
	Location	0.517	0.626		Location	$0.604_{\pm 0.015}$	$0.657_{\pm 0.024}$	$0.517_{\pm 0.037}$	$0.535_{\pm 0.014}$
	Organization	0.500	0.611		Organization	$0.619_{\pm 0.008}$	$0.662_{\pm 0.012}$	$0.516_{\pm 0.012}$	$0.497_{\pm 0.005}$
	Complete	0.593	0.604		Complete	$0.618_{\pm 0.008}$	$0.650_{\pm 0.001}$	$0.517_{\pm 0.002}$	$0.564_{\pm 0.010}$
MM	Text-changed	0.515	0.693	MM	Text-changed	$0.787_{\pm 0.012}$	$0.797_{\pm 0.016}$	$0.499_{\pm 0.002}$	$0.502_{\pm 0.000}$
	Image-changed	0.520	0.661		Image-changed	$0.761_{\pm 0.016}$	$0.821_{\pm 0.012}$	$0.500_{\pm 0.003}$	$0.499_{\pm 0.001}$
TMG	model-generated	0.402	0.551	TMG	model-generated	$0.980_{\pm 0.002}$	$0.984_{\pm 0.001}$	$0.797_{\pm 0.013}$	$0.994_{\pm 0.001}$
	with rumour	0.417	0.563		with rumour	$0.978_{\pm 0.003}$	$0.980_{\pm 0.002}$	$0.798_{\pm 0.009}$	$0.993_{\pm 0.003}$
MTG	model-generated	0.360	0.328	MTG	model-generated	$0.709_{\pm 0.013}$	$0.796_{\pm 0.029}$	$0.956_{\pm 0.012}$	$0.894_{\pm 0.016}$
	with rumour	0.395	0.398		with rumour	$0.705_{\pm 0.016}$	$0.801_{\pm 0.015}$	$0.951_{\pm 0.016}$	$0.891_{\pm 0.008}$
M2N	movie+MM	0.533	0.697	M2N	movie stills+MM	$0.833_{\pm 0.009}$	$0.846_{\pm 0.008}$	$0.521_{\pm 0.004}$	$0.531_{\pm 0.005}$
	movie+MTG	0.421	0.434		movie stills+MTG	$0.712_{\pm 0.007}$	$0.813_{\pm 0.018}$	$0.960_{\pm 0.011}$	$0.902_{\pm 0.010}$

However, the highest accuracies achieved by both models are still moderate, which is expected given the challenging nature of our dataset. The lowest accuracy for both models occurs with MTG. BLIP scores 0.360 for model-generated content, while CLIP scores 0.328. In the zero-shot context, deepfake-generated data, which has not been specifically trained on the models, tends to deceive these models more easily.

In summary, while CLIP generally performs better than BLIP, the moderate accuracies across all types indicate that detecting fabricated content remains challenging in zero-shot situations. This demonstrates the complexity of M^3A and its ability to pose a significant challenge for state-of-the-art models.

Multimodal Misinformation Detection. In multimodal misinformation detection experiments, we extract embeddings separately for each modality using BLIP, CLIP, ViLT, and VisualBERT. For instance, BLIP outputs image and text embeddings, both initially in $\mathbb{R}^{768}$, which are concatenated into $\mathbb{R}^{1536}$. Similarly, CLIP generates embeddings in $\mathbb{R}^{512}$, which are concatenated into $\mathbb{R}^{1024}$. VisualBERT, on the other hand, produces fused embeddings directly. After concatenation, the combined embeddings are passed through several intermediate layers, reducing the dimensionality to $\mathbb{R}^{512}$, followed by ReLU activation and Layer Normalization. Finally, the embeddings are transformed into a one-dimensional output in $\mathbb{R}^1$, serving as the classification output.

For each subtype of the dataset, corresponding to different research questions, we balance the data with a 1:1 ratio of positive (real) and negative (fake) samples. After balancing, the data is split into training, validation, and test sets in a 6:2:2 ratio. The model is trained with a dropout rate of 0.2, runs for 50 epochs with a batch size of 32, a learning rate of 0.001, and uses Binary Cross-Entropy Loss. Early stopping is set to 5 epochs to prevent overfitting, and the model with the lowest validation loss is selected to determine the final test accuracy. This experiment evaluates the model's ability to detect misinformation after fine-tuning.

The results in Table 3(b) indicate that each model has strengths in handling different types of fabricated content, yet the challenging nature of our dataset is evident from the varying performance across methods. CLIP generally outperforms other models. For instance, in the NEM category, CLIP achieves the highest accuracies, such as 0.657 for person-manipulated samples and 0.662 for organization-manipulated samples. For the "NEM-complete" subset, the high accuracy is primarily due to the extensive modifications made to named entities, which result in significant divergence from the original text.

In the MM category, CLIP demonstrates strong performance, particularly in detecting image-altered samples generated by BLIP-2, achieving an accuracy of 0.821. The similar results between the MM and M2N (movie + MM) methods, as well as the MTG and M2N (movie + MTG) methods, are due to their shared approach. The slightly higher performance for M2N suggests that movie content follows a more consistent structure compared to the diverse nature of news content, making it easier for models to capture distinct features.

For TMG and MTG, both models achieve significantly higher accuracies compared to their zero-shot results. CLIP scores 0.984 and BLIP 0.980 for TMG, while ViLT shows surprising strength in MTG, achieving 0.956 for model-generated content and 0.951 for generated samples with rumours. VisualBERT achieves the highest accuracy in TMG, scoring 0.994 for model-generated content. The higher detection results in TMG and MTG, compared to their low zero-shot results, can be largely attributed to the unique textures and patterns that are inherently produced by the generation models used to create these datasets. These textures allow models to achieve high accuracy rates simply by identifying them, rather than understanding the underlying content or context.

The varying performances across different categories and models indicate that our dataset presents a significant challenge, requiring robust and versatile models to effectively detect all types of fabricated content. This underscores the complexity and challenge of M^3A.

4.2. Falsified text–image–audio–video detection

In the misinformed text–image–audio–video sample detection experiments, we use ImageBind and LanguageBind as baselines for these multimodal inputs. Consistent with the training strategy outlined in Section 4.1, we evaluate the models using accuracy, AUC, and AP.

As detailed in Table 4, for NEM, which involves minor textual changes to named entities, M^3A effectively challenges both Imagebind and Languagebind. Imagebind achieves higher accuracies (e.g., 0.829 for the "NEM-complete" type) due to its robust multimodal embedding capabilities that handle textual changes well. In contrast, Languagebind, which relies more on word embeddings, struggles with these tasks, showing accuracies around 0.500. This highlights that M^3A's entity manipulations are sufficiently complex to expose Languagebind's limitations in handling minor textual changes.

For MM, which involves swapping the most similar modalities between samples to create OOC samples, our dataset reveals distinct strengths and weaknesses of chosen models. Both models fail to accurately detect audio-changed samples, with accuracies around 0.590 for Imagebind and 0.582 for Languagebind. Imagebind, focusing on visual information, performs well with text-changed samples (Acc: 0.620) but struggles with visual changes (Acc: 0.498 for both image and video). This difficulty arises because detecting visual mismatches in its own generated embeddings is challenging. Similarly, Languagebind, which emphasizes word embeddings, excels in visual mismatch tasks (Acc: 0.676 for image-changed, 0.689 for video-changed) but struggles with text-changed tasks (Acc: 0.500), indicating it is challenging for Languagebind to detect mismatches using its own generated word embeddings. It is worth noting that, unlike the results obtained using

Table 4
Performance of misinformed text–image–audio–video sample detection for different models. Metrics include accuracy (Acc), area under the curve (AUC), and average precision (AP). MM results are based on features extracted using Imagebind and Languagebind.

Method	Type	Imagebind			Languagebind		
		Acc	AUC	AP	Acc	AUC	AP
NEM	Person	$0.709_{\pm 0.005}$	$0.789_{\pm 0.004}$	$0.778_{\pm 0.004}$	$0.501_{\pm 0.002}$	$0.498_{\pm 0.000}$	$0.497_{\pm 0.001}$
	Location	$0.737_{\pm 0.004}$	$0.812_{\pm 0.004}$	$0.791_{\pm 0.006}$	$0.500_{\pm 0.003}$	$0.498_{\pm 0.001}$	$0.497_{\pm 0.000}$
	Organization	$0.654_{\pm 0.004}$	$0.719_{\pm 0.005}$	$0.703_{\pm 0.004}$	$0.497_{\pm 0.000}$	$0.498_{\pm 0.001}$	$0.497_{\pm 0.000}$
	Complete	$0.829_{\pm 0.001}$	$0.912_{\pm 0.001}$	$0.909_{\pm 0.001}$	$0.500_{\pm 0.001}$	$0.500_{\pm 0.000}$	$0.500_{\pm 0.000}$
MM	Text-changed	$0.620_{\pm 0.003}$	$0.662_{\pm 0.006}$	$0.652_{\pm 0.007}$	$0.500_{\pm 0.001}$	$0.500_{\pm 0.000}$	$0.499_{\pm 0.000}$
	Image-changed	$0.498_{\pm 0.001}$	$0.497_{\pm 0.000}$	$0.497_{\pm 0.001}$	$0.676_{\pm 0.006}$	$0.739_{\pm 0.008}$	$0.710_{\pm 0.012}$
	Audio-changed	$0.590_{\pm 0.002}$	$0.627_{\pm 0.003}$	$0.613_{\pm 0.004}$	$0.582_{\pm 0.005}$	$0.631_{\pm 0.000}$	$0.632_{\pm 0.001}$
	Video-changed	$0.498_{\pm 0.001}$	$0.498_{\pm 0.001}$	$0.497_{\pm 0.002}$	$0.689_{\pm 0.007}$	$0.723_{\pm 0.007}$	$0.702_{\pm 0.006}$
TMG	model-generated	$0.992_{\pm 0.004}$	$1.000_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.996_{\pm 0.000}$	$1.000_{\pm 0.000}$	$1.000_{\pm 0.000}$
	with rumour	$0.994_{\pm 0.002}$	$1.000_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.997_{\pm 0.000}$	$1.000_{\pm 0.000}$	$1.000_{\pm 0.000}$
MTG	model-generated	$0.810_{\pm 0.022}$	$0.899_{\pm 0.011}$	$0.896_{\pm 0.014}$	$0.500_{\pm 0.006}$	$0.500_{\pm 0.000}$	$0.506_{\pm 0.000}$
	with rumour	$0.834_{\pm 0.013}$	$0.920_{\pm 0.001}$	$0.936_{\pm 0.001}$	$0.506_{\pm 0.000}$	$0.500_{\pm 0.000}$	$0.506_{\pm 0.000}$
M2N	movie+MM	$0.625_{\pm 0.004}$	$0.665_{\pm 0.004}$	$0.654_{\pm 0.006}$	$0.501_{\pm 0.001}$	$0.500_{\pm 0.001}$	$0.501_{\pm 0.001}$
	movie+MTG	$0.841_{\pm 0.012}$	$0.911_{\pm 0.015}$	$0.902_{\pm 0.016}$	$0.504_{\pm 0.004}$	$0.501_{\pm 0.001}$	$0.508_{\pm 0.004}$

BLIP-2 in Table 3(b), the MM results in Table 4 are based on features extracted using Imagebind and Languagebind for mismatching. The consistently lower performance of MM in Table 4 further indicates that misinformation generated using MM based on Imagebind and Languagebind presents a greater challenge.

In TMG, both models exhibit nearly perfect performance with accuracies approaching 1. However, as shown in Table 3(a), their zero-shot performance remains low. This suggests that the models have primarily learned to detect specific characteristics of model-generated content, caused by the inherent drawbacks of current generation models, rather than truly understanding the relationships between modalities.

For MTG, Imagebind significantly outperforms Languagebind, achieving accuracies around 0.834, while Languagebind is around 0.500. This highlights that M^3A's multimodality-driven text generation is challenging for models relying on word embeddings.

For M2N, both models perform similarly to their results in MM and MTG. ImageBind outperforms LanguageBind in detecting manipulated multimedia content, particularly in the "movie+MM" type (Acc: 0.625 vs. 0.501), though the performance gap narrows in the "movie+MTG" type. Similar to MM, the detection accuracy for the "movie+MM" type is significantly lower compared to the results in Table 3(b). This further supports the conclusion that samples generated by mismatching based on ImageBind and LanguageBind are highly challenging.

Overall, M^3A effectively exposes the strengths and weaknesses of multimodal models like Imagebind and Languagebind. M^3A's complexity and diversity ensure that even advanced models struggle with certain types of misinformation, highlighting areas for future improvement in multimodal misinformation detection.

4.3. Out-of-distribution (OOD) tests

We perform three types of OOD tests to investigate various factors affecting news content detection.

First, we consider geographical location. As shown in Fig. 5(d), we intentionally select news sources from various regions worldwide. To evaluate the impact of geographical location, we train models on data from 35 North American outlets and evaluate it using data from 6 Asian outlets.

We also examine the impact of themes on classifier performance, since each news article has its own theme (Fig. 6(c)). For text–image pairs, we use news related to the economy as the source and news related to international topics as the target. For text–image–audio–video pairs, we use news related to politics as the source and news concerning the environment as the target.

We then explore the effect of sentiment, as each news article carries its own sentiment (Fig. 6(b)). We conduct sentiment OOD tests for text–image–audio–video pairs by selecting news with negative sentiment as the source and news with positive sentiment as the target.

Additionally, considering the token length restrictions of models like CLIP (77 tokens max) and BLIP (predominantly fewer than 40 tokens), we examine the variability in token lengths within M^3A as mentioned in Fig. 6(a). We set a threshold at 40 tokens, using sources with fewer than 40 tokens and targets with more than 40 tokens, to evaluate the model's adaptability to different content lengths.

We shuffle the source data, splitting it into training and validation sets in a 3:1 ratio, and record the highest validation accuracy. The model achieving this peak validation accuracy is then tested on the target data to observe the test accuracy.

Based on the OOD test results for method NEM with fake text–image sample detection (see Table 5), we can derive that ViLT and VisualBERT perform poorly across all tests, so the analysis focuses on other models as shown below:

- For the geography OOD test, both CLIP and BLIP experience a performance drop when transferring from North American to Asian news outlets, with accuracy dropping from 0.650 to 0.613 for detecting person-manipulated samples. This indicates that geographical variability could impact the performances of chosen models.
- For the theme OOD test, changing the theme from economy to international news leads to a significant drop, with accuracy decreasing from 0.679 to about 0.553 for person-manipulated sample detection. This suggests that thematic changes pose a significant challenge to these models.
- For the text token number OOD test, results are mixed. In detecting text–image samples generated by method NEM, variations in token number did not significantly affect the performance of CLIP and BLIP.

For fake text–image–audio–video sample detection according to method MM (see Table 6), the impact of geography, theme, and text token number showed some differences.

For the geography OOD test, M^3A reveals distinct differences between source and target. When shifting from North American to Asian news outlets, Imagebind shows a slight decrease in accuracy and AUC for text-changed tasks (e.g., Acc: 0.619 to 0.611, AUC: 0.663 to 0.639). Languagebind struggles more with visual information, as seen in the drop in accuracy from 0.682 to 0.636 and AUC from 0.751 to 0.689 in image-changed tasks. This suggests a challenge in adapting to regional variations in textual data and indicates that our geographical manipulations effectively test the models' ability to generalize across different regions.

In sentiment OOD tests, the shift from content with negative motion (source) to content with positive (target) motion significantly impacts performance. For Imagebind, text-changed tasks show a notable drop in accuracy from 0.619 to 0.509 and AUC from 0.663 to

Table 5
OOD analysis of method NEM for falsified text–image sample detection. We report the performance on news outlet geography (source: North America — 35 news outlets, target: Asian — 6 news outlets), news theme (source: Economy, target: International), and text token number (source: <40 tokens, target: >40 tokens). The metric used in this table is accuracy.

Domain	Type	CLIP		BLIP		ViLT		VisualBERT	
		Source	Target	Source	Target	Source	Target	Source	Target
Geography	Person	$0.650_{\pm 0.020}$	$0.613_{\pm 0.056}$	$0.566_{\pm 0.023}$	$0.542_{\pm 0.025}$	$0.500_{\pm 0.031}$	$0.497_{\pm 0.022}$	$0.534_{\pm 0.014}$	$0.538_{\pm 0.018}$
	Location	$0.625_{\pm 0.016}$	$0.611_{\pm 0.022}$	$0.557_{\pm 0.030}$	$0.534_{\pm 0.005}$	$0.513_{\pm 0.009}$	$0.488_{\pm 0.014}$	$0.510_{\pm 0.007}$	$0.519_{\pm 0.004}$
	Organization	$0.662_{\pm 0.006}$	$0.648_{\pm 0.003}$	$0.595_{\pm 0.007}$	$0.566_{\pm 0.011}$	$0.516_{\pm 0.004}$	$0.512_{\pm 0.005}$	$0.526_{\pm 0.037}$	$0.523_{\pm 0.033}$
	Complete	$0.652_{\pm 0.004}$	$0.649_{\pm 0.008}$	$0.609_{\pm 0.003}$	$0.577_{\pm 0.003}$	$0.512_{\pm 0.008}$	$0.504_{\pm 0.001}$	$0.566_{\pm 0.008}$	$0.536_{\pm 0.010}$
Theme	Person	$0.679_{\pm 0.058}$	$0.553_{\pm 0.010}$	$0.558_{\pm 0.041}$	$0.520_{\pm 0.014}$	$0.521_{\pm 0.015}$	$0.506_{\pm 0.004}$	$0.531_{\pm 0.024}$	$0.510_{\pm 0.007}$
	Location	$0.616_{\pm 0.131}$	$0.560_{\pm 0.007}$	$0.573_{\pm 0.037}$	$0.538_{\pm 0.009}$	$0.540_{\pm 0.066}$	$0.492_{\pm 0.003}$	$0.501_{\pm 0.020}$	$0.512_{\pm 0.007}$
	Organization	$0.625_{\pm 0.016}$	$0.580_{\pm 0.011}$	$0.551_{\pm 0.020}$	$0.522_{\pm 0.009}$	$0.518_{\pm 0.029}$	$0.502_{\pm 0.003}$	$0.506_{\pm 0.001}$	$0.507_{\pm 0.005}$
	Complete	$0.611_{\pm 0.025}$	$0.608_{\pm 0.004}$	$0.584_{\pm 0.015}$	$0.542_{\pm 0.004}$	$0.489_{\pm 0.008}$	$0.504_{\pm 0.002}$	$0.537_{\pm 0.027}$	$0.529_{\pm 0.021}$
Length	Person	$0.530_{\pm 0.017}$	$0.564_{\pm 0.016}$	$0.526_{\pm 0.023}$	$0.531_{\pm 0.025}$	$0.507_{\pm 0.011}$	$0.504_{\pm 0.003}$	$0.517_{\pm 0.011}$	$0.510_{\pm 0.006}$
	Location	$0.546_{\pm 0.017}$	$0.557_{\pm 0.017}$	$0.513_{\pm 0.002}$	$0.533_{\pm 0.019}$	$0.470_{\pm 0.018}$	$0.500_{\pm 0.000}$	$0.493_{\pm 0.006}$	$0.510_{\pm 0.007}$
	Organization	$0.597_{\pm 0.030}$	$0.645_{\pm 0.004}$	$0.538_{\pm 0.026}$	$0.552_{\pm 0.025}$	$0.501_{\pm 0.015}$	$0.504_{\pm 0.005}$	$0.532_{\pm 0.024}$	$0.530_{\pm 0.021}$
	Complete	$0.653_{\pm 0.006}$	$0.635_{\pm 0.002}$	$0.631_{\pm 0.013}$	$0.584_{\pm 0.002}$	$0.513_{\pm 0.005}$	$0.503_{\pm 0.004}$	$0.575_{\pm 0.011}$	$0.552_{\pm 0.006}$

Table 6
OOD analysis of method MM for falsified text–image–audio–video sample detection. We report the performance on news outlet geography (source: North America — 35 news outlets, target: Asian — 6 news outlets), news sentiment (source: Negative, target: Positive), news theme (source: Politics, target: Environment), and text token number (source: <40 tokens, target: >40 tokens).

Domain	Type	Imagebind						Languagebind					
		Acc		AUC		AP		Acc		AUC		AP	
		Source	Target	Source	Target	Source	Target	Source	Target	Source	Target	Source	Target
Geography	Text-changed	$0.619_{\pm 0.004}$	$0.611_{\pm 0.002}$	$0.663_{\pm 0.006}$	$0.639_{\pm 0.002}$	$0.654_{\pm 0.004}$	$0.633_{\pm 0.001}$	$0.497_{\pm 0.000}$	$0.500_{\pm 0.000}$	$0.500_{\pm 0.000}$	$0.500_{\pm 0.000}$	$0.503_{\pm 0.000}$	$0.500_{\pm 0.000}$
	Image-changed	$0.499_{\pm 0.001}$	$0.500_{\pm 0.000}$	$0.499_{\pm 0.000}$	$0.500_{\pm 0.000}$	$0.500_{\pm 0.000}$	$0.500_{\pm 0.000}$	$0.682_{\pm 0.015}$	$0.636_{\pm 0.009}$	$0.751_{\pm 0.019}$	$0.689_{\pm 0.001}$	$0.728_{\pm 0.022}$	$0.663_{\pm 0.014}$
	Audio-changed	$0.582_{\pm 0.004}$	$0.602_{\pm 0.001}$	$0.618_{\pm 0.004}$	$0.647_{\pm 0.001}$	$0.612_{\pm 0.004}$	$0.637_{\pm 0.001}$	$0.571_{\pm 0.003}$	$0.563_{\pm 0.003}$	$0.616_{\pm 0.006}$	$0.601_{\pm 0.006}$	$0.616_{\pm 0.005}$	$0.594_{\pm 0.006}$
	Video-changed	$0.499_{\pm 0.001}$	$0.500_{\pm 0.000}$	$0.499_{\pm 0.001}$	$0.500_{\pm 0.000}$	$0.500_{\pm 0.000}$	$0.500_{\pm 0.000}$	$0.646_{\pm 0.002}$	$0.646_{\pm 0.002}$	$0.754_{\pm 0.002}$	$0.706_{\pm 0.002}$	$0.737_{\pm 0.003}$	$0.690_{\pm 0.005}$
Sentiment	Text-changed	$0.619_{\pm 0.008}$	$0.509_{\pm 0.002}$	$0.663_{\pm 0.001}$	$0.498_{\pm 0.003}$	$0.653_{\pm 0.008}$	$0.518_{\pm 0.003}$	$0.501_{\pm 0.000}$	$0.500_{\pm 0.000}$	$0.498_{\pm 0.001}$	$0.500_{\pm 0.000}$	$0.497_{\pm 0.000}$	$0.500_{\pm 0.000}$
	Image-changed	$0.498_{\pm 0.003}$	$0.500_{\pm 0.000}$	$0.496_{\pm 0.002}$	$0.500_{\pm 0.000}$	$0.493_{\pm 0.001}$	$0.500_{\pm 0.000}$	$0.638_{\pm 0.005}$	$0.650_{\pm 0.005}$	$0.690_{\pm 0.006}$	$0.709_{\pm 0.005}$	$0.658_{\pm 0.008}$	$0.684_{\pm 0.005}$
	Audio-changed	$0.566_{\pm 0.004}$	$0.586_{\pm 0.001}$	$0.597_{\pm 0.002}$	$0.623_{\pm 0.001}$	$0.591_{\pm 0.002}$	$0.615_{\pm 0.001}$	$0.571_{\pm 0.003}$	$0.573_{\pm 0.001}$	$0.612_{\pm 0.002}$	$0.617_{\pm 0.002}$	$0.612_{\pm 0.003}$	$0.623_{\pm 0.000}$
	Video-changed	$0.496_{\pm 0.003}$	$0.500_{\pm 0.000}$	$0.495_{\pm 0.002}$	$0.500_{\pm 0.000}$	$0.499_{\pm 0.001}$	$0.500_{\pm 0.000}$	$0.611_{\pm 0.006}$	$0.619_{\pm 0.004}$	$0.663_{\pm 0.003}$	$0.578_{\pm 0.001}$	$0.642_{\pm 0.004}$	$0.663_{\pm 0.003}$
Theme	Text-changed	$0.597_{\pm 0.002}$	$0.557_{\pm 0.008}$	$0.629_{\pm 0.004}$	$0.576_{\pm 0.008}$	$0.624_{\pm 0.004}$	$0.587_{\pm 0.007}$	$0.498_{\pm 0.001}$	$0.500_{\pm 0.000}$	$0.497_{\pm 0.001}$	$0.500_{\pm 0.000}$	$0.500_{\pm 0.000}$	$0.500_{\pm 0.000}$
	Image-changed	$0.495_{\pm 0.003}$	$0.500_{\pm 0.000}$	$0.493_{\pm 0.004}$	$0.500_{\pm 0.000}$	$0.494_{\pm 0.002}$	$0.500_{\pm 0.000}$	$0.628_{\pm 0.004}$	$0.606_{\pm 0.002}$	$0.677_{\pm 0.005}$	$0.656_{\pm 0.003}$	$0.644_{\pm 0.007}$	$0.638_{\pm 0.003}$
	Audio-changed	$0.552_{\pm 0.006}$	$0.572_{\pm 0.000}$	$0.570_{\pm 0.012}$	$0.600_{\pm 0.001}$	$0.566_{\pm 0.011}$	$0.591_{\pm 0.003}$	$0.526_{\pm 0.003}$	$0.555_{\pm 0.006}$	$0.542_{\pm 0.001}$	$0.592_{\pm 0.006}$	$0.540_{\pm 0.002}$	$0.600_{\pm 0.007}$
	Video-changed	$0.494_{\pm 0.001}$	$0.500_{\pm 0.000}$	$0.487_{\pm 0.003}$	$0.500_{\pm 0.000}$	$0.489_{\pm 0.005}$	$0.500_{\pm 0.000}$	$0.604_{\pm 0.012}$	$0.585_{\pm 0.003}$	$0.643_{\pm 0.014}$	$0.628_{\pm 0.003}$	$0.613_{\pm 0.012}$	$0.613_{\pm 0.000}$
Length	Text-changed	$0.626_{\pm 0.003}$	$0.509_{\pm 0.005}$	$0.672_{\pm 0.005}$	$0.476_{\pm 0.011}$	$0.659_{\pm 0.006}$	$0.526_{\pm 0.008}$	$0.500_{\pm 0.000}$	$0.500_{\pm 0.000}$	$0.500_{\pm 0.000}$	$0.500_{\pm 0.000}$	$0.500_{\pm 0.000}$	$0.500_{\pm 0.000}$
	Image-changed	$0.499_{\pm 0.001}$	$0.500_{\pm 0.000}$	$0.499_{\pm 0.001}$	$0.500_{\pm 0.000}$	$0.500_{\pm 0.000}$	$0.500_{\pm 0.000}$	$0.679_{\pm 0.015}$	$0.684_{\pm 0.013}$	$0.747_{\pm 0.018}$	$0.751_{\pm 0.013}$	$0.719_{\pm 0.002}$	$0.730_{\pm 0.012}$
	Audio-changed	$0.592_{\pm 0.002}$	$0.615_{\pm 0.003}$	$0.634_{\pm 0.005}$	$0.664_{\pm 0.004}$	$0.626_{\pm 0.002}$	$0.651_{\pm 0.002}$	$0.575_{\pm 0.003}$	$0.600_{\pm 0.003}$	$0.624_{\pm 0.002}$	$0.666_{\pm 0.003}$	$0.623_{\pm 0.001}$	$0.669_{\pm 0.002}$
	Video-changed	$0.500_{\pm 0.001}$	$0.500_{\pm 0.000}$	$0.498_{\pm 0.001}$	$0.500_{\pm 0.000}$	$0.498_{\pm 0.000}$	$0.500_{\pm 0.000}$	$0.678_{\pm 0.008}$	$0.689_{\pm 0.001}$	$0.746_{\pm 0.010}$	$0.761_{\pm 0.012}$	$0.728_{\pm 0.001}$	$0.748_{\pm 0.013}$

0.498. This indicates that Imagebind models trained on one sentiment perform poorly when tested on another, highlighting generalization issues.

In theme OOD tests, transitioning from "politics" (source) to "environment" (target) impacts model performance. Imagebind's accuracy for text-changed tasks drops from 0.597 to 0.557, with AUC decreasing from 0.629 to 0.576. Similarly, Languagebind shows a decline in audio-changed tasks, where accuracy falls from 0.555 to 0.526 and AUC from 0.592 to 0.542. These results underscore the difficulty in adapting to different thematic content, highlighting the challenge posed by thematic shifts.

In token OOD tests, token length presents a significant challenge for Imagebind when texts lengthen from short (source) to long (target). For text-changed tasks, accuracy decreases from 0.626 to 0.509 and AUC from 0.672 to 0.476, indicating difficulty with increased content complexity. Conversely, Languagebind does not show a significant change, showing a stable performance regardless of text length.

Overall, while the models show some generalization, factors such as geographical location, sentiment shifts, token length, and thematic changes pose challenges. This highlights the need for robust training strategies that accommodate diverse sources, themes, content lengths, and sentiments to enhance real-world model generalization.

4.4. Unimodality detection

In Tables 3(a) and (3b), we present the classification performance on misinformation generated by methods TMG and MTG. The results show that models exhibit low accuracy in zero-shot scenarios but achieve high accuracy after training. This raises the question of whether the models genuinely rely on the relationships between modalities for classification or if they merely learn to recognize the specific patterns in the pre-trained model outputs. To investigate this, we conduct additional unimodality detection experiments.

Table 7 assesses the performance of single-modal (image) misinformation detection. Image features are extracted using Resnet-152, Xception, EfficientNet-B1, and ViT. Resnet-152 (He et al., 2016) is known for its deep residual learning capabilities, Xception (Chollet, 2017) uses depthwise separable convolutions to improve performance, EfficientNet-B1 (Tan and Le, 2019) optimizes accuracy and efficiency, and ViT (Dosovitskiy et al., 2021) leverages transformer architecture for image recognition. These models are well-established for image recognition, each with unique architectures enhancing feature extraction and classification.

The accuracy in Table 7 is consistently high, especially for Xception and ViT, which achieve over 90% accuracy. Considering the results

Table 7
Performance of unimodal (image) misinformation detection. This table shows the accuracy of different models in detecting images generated by the TMG method. "Combined choose" refers to selected outputs from results generated by all three models with the highest similarity to the original text. The metric used in this table is accuracy.

Method	Resnet-152	Xception	EfficientNet-B1	ViT
Glide	0.892	0.989	0.861	0.989
Kandinsky	0.849	0.973	0.846	0.977
SD-Turbo	0.771	0.942	0.812	0.975
Combined choose	0.765	0.941	0.812	0.975

Table 8
Performance of cross-type misinformed text–image–audio–video sample detection. Metrics include accuracy (Acc), area under the curve (AUC), and average precision (AP). The models train on outputs generated by the SD-Turbo type in method TMG and are tested on outputs from both SD-Turbo and Kandinsky types. The training and testing sets do not overlap.

Type	Imagebind			Languagebind		
	Acc	AUC	AP	Acc	AUC	AP
SD-Turbo	1.000	1.000	1.000	1.000	1.000	1.000
Kandinsky	0.551	0.832	0.755	0.538	0.765	0.682

Table 9
Performance of misinformed text–image–audio–video sample detection for "complete" samples and "fixed complete" samples. Metrics include accuracy (Acc), area under the curve (AUC), and average precision (AP).

Type	Imagebind			Languagebind		
	Acc	AUC	AP	Acc	AUC	AP
Complete	0.829	0.912	0.909	0.500	0.500	0.500
Fixed complete	0.753	0.832	0.828	0.500	0.498	0.498

from Tables 3(a) and (3b), we can conclude that the substantial improvement in model performance after fine-tuning stems from model's ability to learn nuanced features within each modality.

4.5. Cross-type detection

Tables 3(b) and 4 show that fine-tuned models on TMG samples achieve near-perfect accuracy, close to 100%, indicating that models effectively capture the characteristics of these generated samples. To assess whether these characteristics improve model generalization, we perform cross-type detection on text–image–audio–video samples.

Table 8 shows that models trained on TMG-SD-Turbo samples, achieve near-perfect accuracy on SD-Turbo outputs. However, performance drops significantly when tested on a different type, such as Kandinsky. This scenario is common in real-world applications, as it is unrealistic to know the generation methods and train a detection model based on the data from the specific generation method in advance.

The low cross-type performance demonstrates that models trained on one TMG type struggle to generalize across other generative methods, highlighting the need to enhance model robustness in detecting diverse types of misinformation.

4.6. Unimodal bias test

As noted by VERITE (Papadopoulos et al., 2024), unimodal bias arises when named entity replacements introduce factual errors, making misinformation easier to detect. For example, replacing "Former German Chancellor Angela Merkel" with "Former German Chancellor David Cameron" leads to a clear factual mistake (see Fig. 4). To further explore the impact of these biases, we conduct additional tests.

We focus on the "NEM-complete" category, as text–image–audio–video samples from this type have high detection accuracy and are most likely to exhibit unimodal biases. To address this, we use the same APIs as in the MTG method to refine the text after named entity replacements. This ensures that the text remains logically consistent and free from obvious errors. Additionally, we provide the original text as a negative sample to ensure the generated output differs from the original, preserving the falsified nature of the content. The most convincing fabricated text, with the highest BLIP cosine similarity to the other modalities in the sample, is selected. The resulting samples are categorized as "fixed complete" and undergo the same tests as the "complete" samples.

The results in Table 9 show that the "fixed complete" samples, where logical inconsistencies are corrected, have lower accuracy, AUC, and AP compared to the untreated "complete" samples from NEM. This suggests that addressing unimodal biases makes misinformation more difficult to detect, underscoring the need to minimize these biases when generating misinformation. These corrected samples are included as an additional component of M^3A, similar to an appendix, to enrich and extend the dataset without replacing the original samples.

Table 10
Performance of rumour detection. Binary classification performance on the test set between "model-generated" and "with rumour" text–image pairs for methods TMG and MTG. The metric used in this table is accuracy.

Method	BLIP	CLIP	ViLT	VisualBERT
TMG	0.498	0.498	0.499	0.500
MTG	0.501	0.501	0.502	0.499

4.7. Rumour detection

The TMG and MTG methods were originally designed for model-generated content detection, but we introduced variations to better reflect real-world distributions. As shown in Fig. 4, by adding rumours into the prompts, TMG and MTG now include both simple generated content and misleading content. In Tables 3(a) and (3b), detection results for 'model-generated' and 'with rumour' are similar, showing low accuracy in zero-shot scenarios but significant improvement after training. To explore this further, we conducted experiments to directly differentiate between these two types using the same models.

In Table 10, the binary classification performance on the test set between "model-generated" and "with rumour" text–image pairs for TMG and MTG methods hovers around 0.500 across all models. This near-random performance indicates that the models struggle to distinguish between samples with or without rumours even after training.

The results support our hypothesis that models rely on the distinctive patterns in model-generated data rather than the consistency across different modalities. This underscores that M^3A presents a real challenge for models, requiring them to go beyond simple pattern recognition.

4.8. Failure case analysis

Fig. 2 introduces the four data types in M^3A: Pristine, Factual Error, OOC Issue, and Model-generated. To further illustrate the challenges, we include Fig. 7, which shows examples of detection failures for each data type.

In Case 1, the original text mentions a firefighter, but the name is replaced with Arnold Schwarzenegger in this sample. People in the image wear masks due to the pandemic, which makes it harder for the model to detect the named entity replacement. In Case 2, the text and image come from different events in 2020 and 2021. Rome's Colosseum closes and reopens multiple times, so the model needs extra information to distinguish between these events. In Case 3, the image is model-generated. Despite the model's strong performance, some high-quality outputs still evade detection. In Case 4, specific textures in

Fig. 7. Examples of failure cases from M^3A. This figure illustrates model detection failures across four data types in M^3A: Factual Error (Case 1), OOC Issue (Case 2), Model-generated (Case 3), and Pristine (Case 4).

pristine samples confuse the model, leading to misidentification as model-generated.

In conclusion, while the model performs well, there is room for improvement. For example, adding more external evidence might further improve detection accuracy.

4.9. Human performance

M^3A maintains a high level of deception in the falsified information. To validate this, we reference the work of NewsCLIPpings (Luo et al., 2021) and conduct a human performance evaluation step.

The experiment proceeds as follows. We randomly select 300 text–image–audio–video pairs from the outputs of methods NEM and MM, totalling 600 pairs. These consist of 150 authentic and 150 falsified examples for each method. We recruit ten volunteers from the University of Queensland. For each pair, five volunteers answer the following three questions without using search engines or large language models. Q1: Do the modalities in this pair match? (1 — yes or 0 — no) Q2: Are you confident in your answer? (1 — yes or 0 — no) Q3: Would you be more confident in your answer if you could use a search engine? (1 — yes or 0 — no).

Based on the results presented in Table 11, the key insights from the database evaluation are as follows: (1) The average accuracy over all samples is 0.697 for NEM and 0.579 for MM. This clearly shows that detecting misinformation in M^3A is challenging for humans, as the accuracy is relatively low. (2) Humans are better at identifying pristine samples than falsified ones, with accuracy of 0.787 and 0.607 in NEM, and 0.723 and 0.436 in MM. This indicates that participants are frequently misled by falsified content, underscoring the database's challenge in creating deceptive and convincing samples. (3) The optimistic accuracy for falsified samples is 0.813 for NEM and 0.713 for MM. This shows that while the task is challenging, a significant portion

Table 11
Human performance result. Metrics include accuracy, optimistic accuracy (at least one participant gives the right answer), mean Q2 value and mean Q3 value. "Correct" and "Wrong" stand for correctly predicted samples and incorrectly predicted samples.

	NEM			MM		
	Overall	Pristine	Falsified	Overall	Pristine	Falsified
Accuracy ↓	0.697	0.787	0.607	0.579	0.723	0.436
Optimistic accuracy ↓	0.877	0.940	0.813	0.820	0.927	0.713
	Overall	Correct	Wrong	Overall	Correct	Wrong
Mean Q2 value ↓	0.639	0.679	0.547	0.614	0.670	0.537
Mean Q3 value ↑	0.681	0.651	0.750	0.684	0.674	0.697

of the falsified samples can still be correctly identified when considering the best-case scenario where at least one participant provides the correct answer. This reflects the inherent complexity and the need for advanced knowledge to navigate the falsified content accurately. (4) Participants show higher confidence in correct predictions, with Q2 scores of 0.679 for NEM and 0.670 for MM, compared to 0.547 and 0.537 for wrong predictions. (5) The mean Q3 scores are higher for incorrect predictions (0.750 for NEM, 0.697 for MM) than for correct ones (0.651 for NEM, 0.674 for MM), suggesting that participants believe additional information would help when they are uncertain or incorrect, highlighting the complexity of M^3A and the potential value of supplementary resources in achieving better results.

The dataset evaluation demonstrates its effectiveness in generating challenging scenarios for misinformation detection. Participants consistently perform worse when identifying falsified content. M^3A successfully captures the complexity of real-world misinformation and contributes to developing more robust detection methods.

5. Conclusion

In this paper, we introduce the Multimedia Misinformation Dataset for Media Authenticity Analysis (M^3A), a comprehensive large-scale dataset designed to address the limitations of existing misinformation detection datasets. By compiling genuine content from 60 prominent news outlets worldwide and generating false content through various techniques, M^3A offers a diverse and extensive collection of over 7 million samples spanning text, images, audio, and video.

Our analysis demonstrates that existing detection models struggle to generalize without specific training, highlighting the complexity and robustness of our dataset. The poor performance of models in zero-shot scenarios and their significant improvement post-training indicate that M^3A effectively challenges models.

By providing multi-class annotations and establishing benchmarks for various misinformation detection tasks, M^3A equips researchers with the tools needed to develop advanced detection models capable of addressing the multifaceted challenges of misinformation. This dataset serves as a valuable resource for advancing research in out-of-context detection, deepfake identification, fact-checking, and out-of-distribution testing, thereby contributing to the development of more robust and generalizable misinformation detection systems.

CRediT authorship contribution statement

Qingzheng Xu: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Huiqiang Chen:** Writing – review & editing, Validation, Investigation, Formal analysis. **Heming Du:** Writing – review & editing, Investigation. **Hu Zhang:** Writing – review & editing, Visualization, Supervision, Investigation. **Szymon Łukasik:** Writing – review & editing, Supervision, Investigation. **Tianqing Zhu:** Writing – review & editing, Supervision, Investigation. **Xin Yu:** Writing – review & editing, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

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

Acknowledgment

This research is supported by the Australian Research Council (ARC) through the ARC Discovery Early Career Researcher Award (DECRA) under Grant DE230100477 and the ARC Discovery Project under Grant DP220100800.

Data availability

Data will be made available on request.

References

Abdelnabi, S., Hasan, R., Fritz, M., 2022a. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In: CVPR.

Abdelnabi, S., Hasan, R., Fritz, M., 2022b. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, IEEE Computer Society, Los Alamitos, CA, USA, pp. 14920–14929. http://dx.doi.org/10.1109/CVPR52688.2022.01452, URL: https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01452.

Biamby, G., Luo, G., Darrell, T., Rohrbach, A., 2022. Twitter-COMMs: Detecting climate, COVID, and military multimodal misinformation. In: Carpuat, M., de Marneffe, M.-C., Meza Ruiz, I.V. (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Seattle, United States, pp. 1530–1549. http://dx.doi.org/10.18653/v1/2022.naacl-main.110, URL: https://aclanthology.org/2022.naacl-main.110.

Boididou, C., Andreadou, K., Papadopoulos, S., Dang Nguyen, D.T., Boato, G., Riegler, M., Kompatsiaris, Y., et al., 2015. Verifying multimedia use at mediaeval 2015. In: MediaEval 2015. Vol. 1436, CEUR-WS.

Boididou, C., Papadopoulos, S., Dang Nguyen, D.T., Boato, G., Riegler, M., Petlund, A., Kompatsiaris, I., 2016. Verifying multimedia use at MediaEval 2016.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D., 2020. Language models are few-shot learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (Eds.), Advances in Neural Information Processing Systems. Vol. 33, Curran Associates, Inc., pp. 1877–1901, URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Chen, R., Chen, X., Ni, B., Ge, Y., 2020. SimSwap: An efficient framework for high fidelity face swapping. In: ACM MM.

Chien, C.-M., Lin, J.-H., Huang, C.-y., Hsu, P.-c., Lee, H.-y., 2021. Investigating on incorporating pretrained and learnable speaker representations for multi-speaker multi-style text-to-speech. In: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 8588–8592. http://dx.doi.org/10.1109/ICASSP39728.2021.9413880.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions. In: CVPR. pp. 1251–1258.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In: North American Chapter of the Association for Computational Linguistics. URL: https://api.semanticscholar.org/CorpusID:52967399.

Dolhansky, B., Howes, R., Pflaum, B., Baram, N., Ferrer, C.C., 2019. The deepfake detection challenge (DFDC) preview dataset. arXiv preprint arXiv:1910.08854.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations. URL: https://openreview.net/forum?id=YicbFdNTTy.

Dufour, N., Pathak, A., Samangouei, P., Hariri, N., Deshetti, S., Dudfield, A., Guess, C., Escayola, P.H., Tran, B., Babakar, M., Bregler, C., 2024. AMMeBa: A large-scale survey and dataset of media-based misinformation in-the-wild. arxiv:2405.11697, URL: https://arxiv.org/abs/2405.11697.

Elizalde, B., Deshmukh, S., Al Ismail, M., Wang, H., 2023a. Clap learning audio concepts from natural language supervision. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 1–5.

Elizalde, B., Deshmukh, S., Wang, H., 2023b. Natural language supervision for general-purpose audio representations. arxiv:2309.05767. URL: https://arxiv.org/abs/2309.05767.

Gao, G., Huang, H., Fu, C., Li, Z., He, R., 2021. Information bottleneck disentanglement for identity swapping. In: CVPR.

Ghosal, D., Majumder, N., Mehrish, A., Poria, S., 2023. Text-to-audio generation using instruction guided latent diffusion model. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 3590–3598.

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I., 2023. ImageBind: One embedding space to bind them all. In: CVPR.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. Adv. Neural Inf. Process. Syst. 27.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR.

Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A., 2020. spaCy: Industrial-strength Natural Language Processing in Python. http://dx.doi.org/10.5281/zenodo.1212303.

Huang, J., Wang, X., Du, B., Du, P., Xu, C., 2021. Deepfake MNIST+: A deepfake facial animation dataset. In: ICCV. pp. 1973–1982.

Jaiswal, A., Sabir, E., AbdAlmageed, W., Natarajan, P., 2017. Multimedia semantic integrity assessment using joint embedding of images and text. In: Proceedings of the ACM International Conference on Multimedia. MM.

Jiang, L., Li, R., Wu, W., Qian, C., Loy, C.C., 2020. DeeperForensics1.0: A large-scale dataset for real-world face forgery detection. In: CVPR.

Jin, Z., Cao, J., Guo, H., Zhang, Y., Luo, J., 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In: Proceedings of the 25th ACM International Conference on Multimedia. MM '17, Association for Computing Machinery, New York, NY, USA, pp. 795–816. http://dx.doi.org/10.1145/3123266.3123454.

Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H., 2023. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. arXiv preprint arXiv:2303.13439.

Khattar, D., Goud, J.S., Gupta, M., Varma, V., 2019. MVAE: Multimodal variational autoencoder for fake news detection. In: The World Wide Web Conference. WWW '19, Association for Computing Machinery, New York, NY, USA, pp. 2915–2921. http://dx.doi.org/10.1145/3308558.3313552.

Kim, J., Kong, J., Son, J., 2021a. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In: International Conference on Machine Learning. PMLR, pp. 5530–5540.

Kim, W., Son, B., Kim, I., 2021b. Vilt: Vision-and-language transformer without convolution or region supervision. In: International Conference on Machine Learning. PMLR, pp. 5583–5594.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L., 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp. 7871–7880. http://dx.doi.org/10.18653/v1/2020.acl-main.703, URL: https://aclanthology.org/2020.acl-main.703.

Li, J., Li, D., Savarese, S., Hoi, S., 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv:2301.12597.

Li, J., Li, D., Xiong, C., Hoi, S., 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. PMLR, pp. 12888–12900.

Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S., 2020. Celeb-DF: A large-scale challenging dataset for deepfake forensics. In: CVPR.

Li, L.H., Yatskar, M., Yin, D., Hsieh, C.-J., Chang, K.-W., 2019. VisualBERT: A simple and performant baseline for vision and language. arXiv:1908.03557.

Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W., Plumbley, M.D., 2023. AudioLDM: Text-to-audio generation with latent diffusion models. In: Proceedings of the International Conference on Machine Learning.

Liu, F., Wang, Y., Wang, T., Ordonez, V., 2021. Visual news: Benchmark and challenges in news image captioning. In: EMNLP.

Loureiro, D., Barbieri, F., Neves, L., Espinosa Anke, L., Camacho-collados, J., 2022. TimeLMs: Diachronic language models from Twitter. In: Basile, V., Kozareva, Z., Stajner, S. (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Association for Computational Linguistics, Dublin, Ireland, pp. 251–260. http://dx.doi.org/10.18653/v1/2022.acl-demo.25, URL: https://aclanthology.org/2022.acl-demo.25.

Luo, G., Darrell, T., Rohrbach, A., 2021. NewsCLIPpings: Automatic generation of out-of-context multimodal media. In: EMNLP.

Mishra, S., Suryavardan, S., Bhaskar, A., Chopra, P., Reganti, A.N., Patwa, P., Das, A., Chakraborty, T., Sheth, A.P., Ekbal, A., et al., 2022. FACTIFY: A multi-modal fact verification dataset. In: DE-FACTIFY@ AAAI.

Müller-Budack, E., Theiner, J., Diering, S., Idahl, M., Ewerth, R., 2020. Multimodal analytics for real-world news using measures of cross-modal entity consistency. In: ACM ICMR.

Nagrani, A., Chung, J.S., Xie, W., Zisserman, A., 2020. Voxceleb: Large-scale speaker verification in the wild. Comput. Speech Lang. 60, 101027.

Nakamura, K., Levy, S., Wang, W.Y., 2019. R/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In: Proceedings of the International Conference on Language Resources and Evaluation. LREC.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M., 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741.

Nielsen, D.S., McConville, R., 2022. Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '22, Association for Computing Machinery, New York, NY, USA, pp. 3141–3153. http://dx.doi.org/10.1145/3477495.3531744.

Papadopoulos, S.-I., Koutlis, C., Papadopoulos, S., Petrantonakis, P., 2023. Synthetic misinformers: Generating and combating multimodal misinformation. In: Proceedings of the 2nd ACM International Workshop on Multimedia AI Against Disinformation. MAD '23, Association for Computing Machinery, New York, NY, USA, pp. 36–44. http://dx.doi.org/10.1145/3592572.3592842.

Papadopoulos, S.-I., Koutlis, C., Papadopoulos, S., Petrantonakis, P.C., 2024. VERITE: a robust benchmark for multimodal misinformation detection accounting for unimodal bias. Int. J. Multimedia Inf. Retr. 13 (1), 4.

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

Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I., 2023. Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning. PMLR, pp. 28492–28518.

Razzhigaev, A., Shakhmatov, A., Maltseva, A., Arkhipkin, V., Pavlov, I., Ryabov, I., Kuts, A., Panchenko, A., Kuznetsov, A., Dimitrov, D., 2023. Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion. arXiv:2310.03502.

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M., 2019. Faceforensics++: Learning to detect manipulated facial images. In: ICCV.

Rui Shao, T.W., Liu, Z., 2023. Detecting and grounding multi-modal media manipulation. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 6904–6913.

Sabir, E., AbdAlmageed, W., Wu, Y., Natarajan, P., 2018. Deep multimodal image repurposing detection. In: ACM MM.

Sauer, A., Lorenz, D., Blattmann, A., Rombach, R., 2023. Adversarial diffusion distillation. arXiv preprint arXiv:2311.17042.

Shivangi Aneja, C.B., Nießner, M., 2023. COSMOS: Catching out-of-context image misuse using self-supervised learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 37, pp. 14084–14092.

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H., 2020. FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big Data 8 (3), 171–188.

Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. PMLR, pp. 6105–6114.

Tan, R., Saenko, K., Plummer, B.A., 2020. Detecting cross-modal inconsistency to defend against neural fake news. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. EMNLP.

Thies, J., Zollhöfer, M., ner, M.N., 2019. Deferred neural rendering: Image synthesis using neural textures. ACM Trans. Graph. 38 (4), 1–12.

Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., Nießner, M., 2016. Face2Face: Real-time face capture and reenactment of RGB videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2387–2395.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T., 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288.

Wang, B., Komatsuzaki, A., 2021. GPT-J-6B: A 6 billion parameter autoregressive language model. https://github.com/kingoflolz/mesh-transformer-jax.

Wang, S., Li, L., Ding, Y., Fan, C., Yu, X., 2021. Audio2head: audio-driven one-shot talking-head generation with natural head motion. In: Proceedings of the 30th International Joint Conference on Artificial Intelligence. IJCAI.

Wang, S., Li, L., Ding, Y., Yu, X., 2022. One-shot talking face generation from single-speaker audio-visual correlation learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. 36, (3), AAAI, pp. 2531–2539.

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., Gao, J., 2018. EANN: Event adversarial neural networks for multi-modal fake news detection. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD '18, Association for Computing Machinery, New York, NY, USA, pp. 849–857. http://dx.doi.org/10.1145/3219819.3219903.

Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S., 2023. Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571.

Wu, Y., Meng, Y., Hu, Z., Li, L., Wu, H., Zhou, K., Xu, W., Yu, X., 2024. Text-guided 3d face synthesis-from generation to editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 1260–1269.

Xu, Q., Du, H., Chen, H., Liu, B., Yu, X., 2024. Mmooc: a multimodal misinformation dataset for out-of-context news analysis. In: Australasian Conference on Information Security and Privacy. ACISP, Springer, pp. 444–459.

Yang, X., Li, Y., Lyu, S., 2019. Exposing deep fakes using inconsistent head poses. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 8261–8265.

Zhou, T., Wang, W., Liang, Z., Shen, J., 2021. Face forensics in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5778–5788.

Zhu, B., Lin, B., Ning, M., Yan, Y., Cui, J., HongFa, W., Pang, Y., Jiang, W., Zhang, J., Li, Z., Zhang, C.W., Li, Z., Liu, W., Yuan, L., 2024. LanguageBind: Extending video-language pretraining to N-modality by language-based semantic alignment. In: The Twelfth International Conference on Learning Representations. URL: https://openreview.net/forum?id=QmZKc7UZCy.