MULTIBENCH++: A Unified and Comprehensive Multimodal Fusion Benchmarking Across Specialized Domains

Leyan Xue[1], Changqing Zhang[1*], Kecheng Xue[2], Xiaohong Liu[3], Guangyu Wang[2], Zongbo Han[2*]

[1]School of Artificial Intelligence, Tianjin University
[2]State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications
[3]Institute of Medical Artificial Intelligence, South China Hospital, Medical School, Shenzhen University
{xueleyan, zhangchangqing}@tju.edu.cn, xhliu17@gmail.com, {kecheng.xue, guangyu.wang, zongbo}@bupt.edu.cn

Abstract

Although multimodal fusion has made significant progress, its advancement is severely hindered by the lack of adequate evaluation benchmarks. Current fusion methods are typically evaluated on a small selection of public datasets, a limited scope that inadequately represents the complexity and diversity of real-world scenarios, potentially leading to biased evaluations. This issue presents a twofold challenge. On one hand, models may overfit to the biases of specific datasets, hindering their generalization to broader practical applications. On the other hand, the absence of a unified evaluation standard makes fair and objective comparisons between different fusion methods difficult. Consequently, a truly universal and high-performance fusion model has yet to emerge. To address these challenges, we have developed a large-scale, domain-adaptive benchmark for multimodal evaluation. This benchmark integrates over 30 datasets, encompassing 15 modalities and 20 predictive tasks across key application domains. To complement this, we have also developed an open-source, unified, and automated evaluation pipeline that includes standardized implementations of state-of-the-art models and diverse fusion paradigms. Leveraging this platform, we have conducted large-scale experiments, successfully establishing new performance baselines across multiple tasks. This work provides the academic community with a crucial platform for rigorous and reproducible assessment of multimodal models, aiming to propel the field of multimodal artificial intelligence to new heights.

Code — https://github.com/ravexly/MultiBenchplus
Extended version — https://arxiv.org/abs/2511.06452

1 Introduction

Multimodal data, such as text, images, and sensor signals, is driving the next generation of artificial intelligence. Through a technique known as Multimodal Fusion, AI systems can integrate and understand information from these diverse sources, achieving a more comprehensive, accurate, and robust understanding than is possible with any single source (Baltrušaitis, Ahuja, and Morency 2018; Xu, Zhu, and Clifton 2023). This capability is a key driver for advancing AI to higher levels of intelligence and shows immense potential in fields like autonomous driving and medical diagnostics (Caesar et al. 2020; Azam et al. 2022).

However, a significant divergence exists between multimodal research and other domains. While fields such as natural language processing and graph learning have successfully converged on dominant architectural paradigms, specifically the Transformer and GNNs, the multimodal domain conspicuously lacks an equivalent unified, foundational framework. Progress remains highly fragmented; researchers typically validate new methods on a small, bespoke selection of classic datasets. This reliance on siloed benchmarks is a key bottleneck. It not only leads to models overfitting to specific data biases and prevents fair objective comparisons, but more importantly, it has hindered the systematic search for a truly general-purpose fusion architecture. Four years ago, the groundbreaking MULTIBENCH (Liang et al. 2021) framework partially addressed this by providing a unified evaluation platform. Yet, with the field's rapid evolution, its limitations are now apparent, and it is no longer sufficient to meet today's challenges.

The urgent need for a new foundational platform stems from two primary trends. The first is the explosive growth in data combinatorial complexity. Unlike unimodal tasks, real-world applications in medical imaging, IoT, and autonomous driving (Hu et al. 2023; Kong et al. 2011) span a vast, heterogeneous spectrum. This combinatorial effect, where different data combinations can yield entirely different analytical conclusions, means older, simpler benchmarks can no longer demonstrate a model's robustness or adaptability to real-world complexity. The second trend is the rapid evolution of fusion models, especially Transformer-based methods (Wei et al. 2020; Wang et al. 2022; Xu, Zhu, and Clifton 2023). Without a standardized, complex testbed, it is impossible to determine if these new models are truly general-purpose or simply adept at specific data combinations. Therefore, a platform that forces models to confront this combinatorial complexity has become essential to guide the search for a foundational architecture.

To address key challenges in evaluating next-generation multimodal fusion models, we introduce MULTIBENCH++, a new, large-scale benchmark. Instead of being a simple incremental update, MULTIBENCH++ represents a significant leap forward in terms of scale, domain diversity, and suitability for modern architectures.

*Corresponding author.

Domain	Dataset	Modalities	# Samples	Prediction Task
Remote Sensing	Houston2013	$\{h, L\}$	14,999	Land Cover Classification
	Houston2018	$\{h, L\}$	2,018,910	Land Cover Classification
	MUUFL Gulfport	$\{h, L\}$	53,687	Land Cover Classification
	Trento	$\{h, L\}$	30,214	Land Cover Classification
	Berlin	$\{h, s\}$	464,671	Land Cover Classification
	MDAS (Augsburg)	$\{h, s\}$	78,294	Land Cover Classification
	ForestNet	$\{c, i, m\}$	2,757	Forest Type Mapping
Medical AI	TCGA-BRCA	$\{o_1, o_2, o_3\}$	875	Survival Prediction, Subtype Classification
	ROSMAP	$\{o_1, o_2, o_3\}$	351	Disease Progression Prediction
	SIIM-ISIC	$\{i, M\}$	33,126	Malignant Tumor Classification
	Derm7pt	$\{i, M\}$	1,011	Lesion Diagnosis Prediction
	GAMMA	$\{f, O\}$	100	Glaucoma Grading
	MIMIC-III	$\{c, t\}$	36,212	Mortality Prediction
	MIMIC-CXR	$\{c, t\}$	372,147	Mortality Prediction, Multilabel Classification
	eICU	$\{c, t\}$	7,637	Mortality Prediction
	TCGA	$\{o_1, o_2, o_3\}$	306	Subtype classification, Tumor Malignancy Grading
Affective Computing &Social Media Understanding	MELD	$\{a, t\}$	13,708	Emotion/Sentiment Recognition
	IEMOCAP	$\{a, v, t\}$	7,433	Emotion/Sentiment Recognition
	MAMI	$\{i, t\}$	11,000	Misogyny Content Detection
	Memotion	$\{i, t\}$	6,831	Offensive Content Detection
	MUTE	$\{i, t\}$	4,156	Hate Speech Detection
	MultiOFF	$\{i, t\}$	743	Offensive Content Detection
	MET-Meme(C)	$\{i, t\}$	2,299	Metaphor/Emotion/Intent Recognition
	MET-Meme(E)	$\{i, t\}$	1,053	Metaphor/Emotion/Intent Recognition
	CH-SIMS	$\{a, v, t\}$	2,281	Sentiment Analysis
	CH-SIMS v2.0	$\{a, v, t\}$	4,403	Sentiment Analysis
	Twitter2015	$\{i, t\}$	5,338	Multimodal Named Entity Recognition
	Twitter1517	$\{i, t\}$	4,672	Multimodal Named Entity Recognition
Others	MIRFLICKR	$\{i, t\}$	20,015	Image Retrieval
	CUB Image-Caption	$\{i, t\}$	117,880	Fine-grained Classification
	SUN-RGBD	$\{i, d\}$	9,504	Scene Understanding, Object Detection
	NYUDv2	$\{i, d\}$	1,863	Scene Understanding, Object Detection
	UPMC-Food101	$\{i, t\}$	90,686	Food Recognition
	MVSA-Single	$\{i, t\}$	2,592	Sentiment Analysis
	MNIST-SVHN	$\{i, i\}$	660,680	Digit Recognition
	N-MNIST+N-TIDIGITS	$\{e, i\}$	4,050	Digit Recognition
	E-MNIST+EEG	$\{i, k\}$	702	Digit Recognition

Table 1: MULTIBENCH++ offers a unified benchmark suite of 37 multimodal datasets spanning a wide spectrum of research fields, data scales, input modalities(a: audio, c: clinical/tabular, d: depth/DSM, e: events/spiking, f: 2D fundus, g: GIS, h: HSI, i: image, k: time-series, L: LiDAR, m: multispectral, M: metadata, o: multi-omics, O: 3D OCT, s: SAR, t: text, v: video, with omics sub-typed as o_1: mRNA, o_2: miRNA, o_3: DNA), and downstream tasks.

Its core contributions are threefold:

- Expanded scale and domain coverage. MULTIBENCH++ brings together over 30 datasets, more than doubling the size of its predecessor. More importantly, it extends into highly complex and specialized domains, including Remote Sensing, Healthcare, Affective Computing, and Social Media Analysis. These domains present unique data fusion challenges.
- Designed for rigorous testing of advanced architectures. Its datasets are carefully chosen for their high complexity, rich interplay between modalities, and naturally occurring missing data, creating a challenging test environment. This design allows for rigorous testing of advanced, Transformer-based architectures and novel fusion techniques.
- An open-source framework for robust and reproducible evaluation. To ensure fair and rigorous scientific comparisons, MULTIBENCH++ includes a standardized, open-source evaluation framework. This framework provides standardized data splits, Robustness Probes, and a set of strong baseline models that have been carefully tuned using Automated Hyperparameter Optimization. This infrastructure is designed to lower the barrier to entry for researchers and ensure that future innovations can be reliably evaluated on a fair and consistent foundation.

2 Related Works

Comparisons with Related Benchmarks. Multimodal research has been driven by a series of influential benchmarks. Foundational datasets for visual question answer-

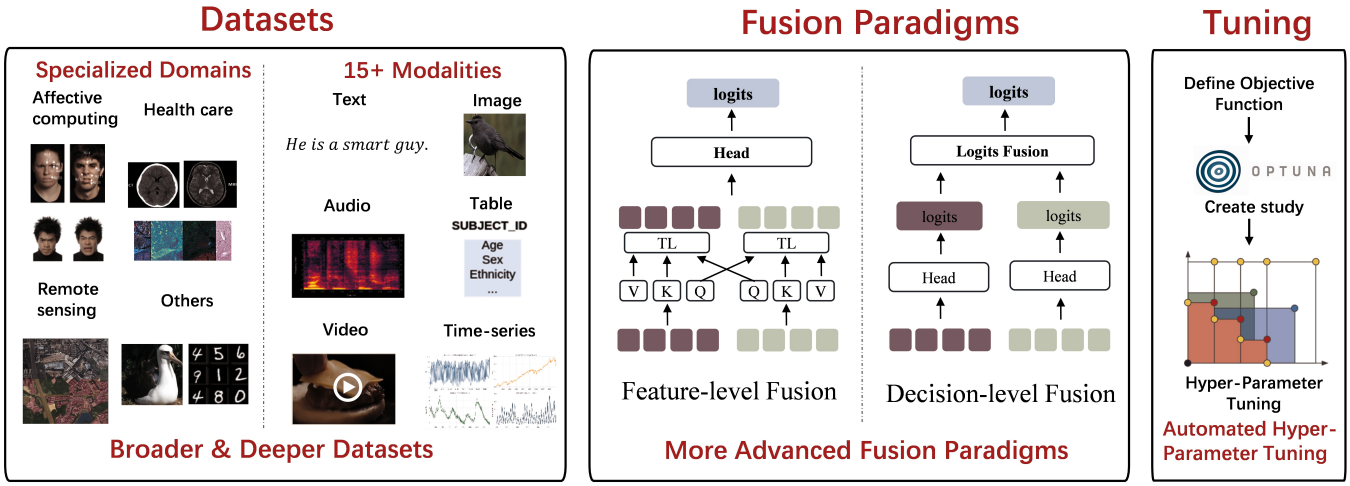

Figure 1: An overview of the MULTIBENCH++ framework, highlighting our core contributions. (Left) We introduce a broader and deeper collection of datasets, significantly expanding into more specialized domains and data modalities. (Center) We integrate more advanced fusion paradigms, including feature-level transformer-based fusion and decision-level fusion. (Right) We provide an automated hyper-parameter tuning platform, powered by Optuna, to ensure robust and reproducible evaluation.

ing, such as VQA-v2 (Goyal et al. 2017), established large-scale, open-ended visual reasoning as a core challenge. In parallel, benchmarks for multimodal sentiment analysis and emotion recognition, like CMU-MOSI (Zadeh et al. 2016) and the larger CMU-MOSEI (Zadeh et al. 2018), provided key testbeds for integrating language, visual, and acoustic signals. To standardize evaluation across this growing landscape, frameworks like MULTIBENCH (Liang et al. 2021) were introduced to offer a unified, reproducible testbed for assessing model robustness across a diverse set of tasks. Building on this principle, our work expands this suite with 20 additional datasets, contributing to a broader trend of comprehensive evaluation that also includes integrating more competitive methods and developing automated tuning platforms.

Other works introduce new domains, such as MM-GRAPH (Zhu et al. 2025), which integrates visual features into graph-based tasks, and Dyn-VQA (Li et al. 2024), which tests dynamic question answering requiring multi-hop retrieval.

Multimodal Fusion. The core challenge in multimodal learning is fusion: the effective combination of information from different modalities. Classical approaches are categorized by the architectural stage at which fusion occurs: early (feature-level) fusion concatenates raw or low-level features (Ramachandram and Taylor 2017; Atrey et al. 2010), while late (decision-level) fusion combines outputs from modality-specific models (Soleymani et al. 2017; Han et al. 2022; Zhang et al. 2023). The advent of the Transformer has made attention-based fusion the dominant paradigm (Xu, Zhu, and Clifton 2023). These methods can be broadly classified as single-stream, where multimodal inputs are concatenated and processed by a unified encoder, or multi-stream, which uses separate encoders for each modality followed by cross-attention mechanisms to integrate information (Nagrani et al. 2021). Such architectures allow for nuanced, dynamically-weighted integration of modal features. More recent research also focuses on developing fusion techniques that are robust to real-world challenges like noisy, incomplete, or imbalanced data (Zhang et al. 2024).

3 MULTIBENCH++ : A Broader & Deeper Multimodal Benchmark

3.1 Background

Multimodal datasets differ in both the types of information they provide and the ways that information is encoded. Remote sensing archives couple spectral bands with LiDAR point clouds, while electronic health records weave free-text notes together with structured lab values and pathology slides. Affective repositories, in turn, align video frames to audio streams and wearable signals. These collections are rarely designed for joint use, so their formats, resolutions, and noise levels diverge. A benchmark must therefore expose models to this heterogeneity. As shown in Fig. 1, MULTIBENCH++ addresses the current evaluation gap by collecting over thirty datasets drawn from highly specialized domains including remote sensing, healthcare, affective computing and social media. These specialized domains are not arbitrary. They represent critical frontiers where multimodal integration is pivotal for scientific and societal progress. The inclusion of such a wide array of data sources ensures that the benchmark rigorously tests a model's ability to generalize across fundamentally different data structures and noise profiles, moving beyond single-domain evaluations.

3.2 Datasets

Remote Sensing for Environmental Intelligence The remote-sensing domain for environmental intelligence tack-

Algorithm 1: End-to-End Workflow with Optuna.

```
# 1. Load data and define models
train_loader, val_loader, test_loader, n_classes = get_loader()
encoders = [Modality1Encoder(args), Modality2Encoder(args)]
fusion = TMC(n_classes)
head = get_head(n_classes, decision=True)
# 2. Define a minimal objective function
def objective(trial):
    # Suggest hyperparameters
    params = {'lr': trial.suggest_loguniform('lr', 1e-5, 1e-3), ...}
    # Train a model and return its validation accuracy
    val_accuracy = train(encoders, fusion, head, params, ...)
    return val_accuracy
# 3. Run the hyperparameter search
study = optuna.create_study(direction='maximize')
study.optimize(objective, n_trials=10)
# 4. Get the best parameters and build the final model
final_model = build_model(study.best_trial.number)
# 5. Evaluate the final, optimized model
test_accuracy = test(final_model, test_loader)
```

les land-cover classification, target detection, spectral unmixing, and related tasks by fusing data whose physical origins differ fundamentally. Optical and hyperspectral systems capture surface chemistry yet remain weather-dependent; SAR measures microwave backscatter day-and-night; LiDAR delivers centimetre-level topography and canopy structure. Integrating these streams requires reconciling disparate spatial resolutions, geometries, and noise statistics. Foundational datasets establish this task-oriented landscape. Houston2013 and Houston2018 (Debes et al. 2014; Xu et al. 2018) combine hyperspectral imagery with LiDAR for urban land-cover classification. MUUFL Gulfport (Gader et al. 2013) adds co-registered hyperspectral and LiDAR data over a university campus for classification and rare-target detection. Trento (University of Trento 2022) provides a rural counterpart with hyperspectral and LiDAR. Berlin (Okujeni, van der Linden, and Hostert 2016) fuses PolSAR and hyperspectral data, while ForestNet (Irvin et al. 2020) couples satellite imagery and airborne LiDAR for forest-type and biomass mapping. The recent MDAS dataset (Hu et al. 2023) enriches the benchmark suite with simultaneous SAR, multispectral, hyperspectral, DSM, and GIS layers, supporting resolution enhancement, unmixing, and classification. Together, these datasets constitute a systematic test-bed for advancing theoretically grounded and practically robust environmental-intelligence algorithms.

Medical AI for Diagnostics and Prognosis The medical-intelligence domain addresses survival prediction, malignancy classification, disease-progression modelling and related tasks by fusing exceptionally heterogeneous data streams. Gigapixel whole-slide images quantify tissue morphology; high-dimensional omics profiles capture molecular aberrations; dense time-series vital signs and concise clinical narratives encode patient trajectories. Integrating these modalities demands reconciling extreme differences in resolution, scale and noise, while preserving clinical interpretability. Benchmark datasets anchor this landscape. TCGA-BRCA (Weinstein et al. 2013) couples WSIs with multi-omics for breast-cancer survival and subtype analysis. ROSMAP (Bennett et al. 2018) supplies longitudinal multi-omics and pathology to chart Alzheimer progression. In dermatology, SIIM-ISIC (Rotemberg et al. 2021) and Derm7pt (Kawahara et al. 2018) pair dermoscopic images with patient metadata for melanoma detection. GAMMA (Wu et al. 2023) fuses 2D fundus photographs with 3D OCT volumes for glaucoma grading and optic-disc/cup segmentation. MIMIC-III (Johnson et al. 2016), MIMIC-CXR (Johnson et al. 2019) and eICU (Pollard et al. 2018) deliver large-scale ICU time series merged with static clinical records to support mortality prediction and disease-code classification.

Affective Computing and Social Media Understanding The Affective Computing domain addresses emotion recognition, sarcasm detection, sentiment analysis and related high-level tasks by fusing text, acoustic, visual and cultural cues that are frequently incongruent or metaphorical. Dyadic and multi-party conversations introduce temporal alignment challenges, while internet memes overlay visual symbols with rapidly shifting socio-cultural contexts. Effective integration demands models that can resolve cross-modal sarcasm, capture long-range conversational flow and remain sensitive to cultural nuance. Benchmark datasets collectively span these phenomena. MELD (Poria et al. 2019) and IEMOCAP (Busso et al. 2008) provide temporally aligned audio, video and transcriptions for multi-party and dyadic emotion recognition. MUTE (Hossain, Sharif, and Hoque 2022) extends the conversational setting to multilingual scenarios. MAMI (Fersini et al. 2022), MultiOFF (Suryawanshi et al. 2020), Memotion (Sharma et al. 2020) and MET-Meme (Xu et al. 2022) (Chinese & English versions) jointly encode images and text for the detection of misogynistic, offensive and metaphorical content in memes. CH-SIMS (Yu et al. 2020) and its successor CH-SIMS v2.0 (Liu et al. 2022) deliver fine-grained Chinese multimodal sentiment annotations with explicit modality-importance scores. Twitter2015 and

Twitter1517 (Zhang et al. 2018; Lu et al. 2018; Chen et al. 2023) close the loop with classic social-media tasks, linking text and images for named-entity recognition and sentiment polarity prediction.

Others This domain addresses image–text retrieval, fine-grained classification, digit recognition, scene understanding and related tasks by fusing heterogeneous yet tightly aligned modalities. Integrating vision with language, depth, audio or event streams demands reconciling distinct resolutions, sampling rates and noise distributions while preserving interpretability. Benchmark datasets anchor this landscape. MIRFLICKR-25K (Huiskes and Lew 2008) couples images with user tags for large-scale retrieval. MVSA-Single (Niu et al. 2016) supplies tweet images and text for visual–sentiment classification. CUB Image-Caption (Shi et al. 2019) pairs bird photographs with textual descriptions for fine-grained classification, while MNIST-SVHN (Shi et al. 2019) aligns handwritten and street-view digits across domains. SUN-RGBD (Song, Lichtenberg, and Xiao 2015) and NYUDv2 (Silberman et al. 2012) provide RGB–depth pairs for scene understanding and object detection, and UPMC-Food101 (Wang et al. 2015) fuses food images with recipe text for cross-modal recognition. Following Lin et al. (2025), we further combine N-MNIST+N-TIDIGITS (Orchard et al. 2015; Anumula et al. 2018) to synchronise frame-based and event-based vision with spoken digits, and E-MNIST+EEG (Cohen et al. 2017; Willett et al. 2021) to link character images with electroencephalography signals for cognitive-state-aware digit recognition.

3.3 Evaluation Protocol

We follow MULTIBENCH's holistic evaluation with only minor adjustments. For every dataset and method, we report performance on the test fold using task-specific metrics (e.g., accuracy, macro-F1, AUPRC, or MSE). Each run is performed 3 times using different random seeds, all under the same hardware configuration.

4 MULTIBENCH++ Algorithms: More Advanced Fusion Paradigms

We introduce a complete, end-to-end framework for systematic multimodal evaluation. The framework is built around two classes of fusion methodologies: four Transformer-centric paradigms to model complex cross-modal interactions, and two modules for efficient decision-level logits fusion. To enable robust and reproducible experimentation, we further develop an automated hyperparameter optimization engine based on Optuna (Akiba et al. 2019). This engine facilitates a systematic exploration and optimization, allowing for an efficient identification of optimal configurations.

4.1 Transformer-Based Feature Fusion Architectures

We evaluate several Transformer-based architectures for multimodal feature fusion. Each model is designed to accept a set of modality-specific input tensors $\{x_i\}_{i=1}^k$ and produce a unified representation vector g.

Hierarchical Attention (Multi-to-One) This model first encodes each modality independently using a shallow, modality-specific Transformer encoder $\mathcal{E}_i$. The resulting classification tokens are then concatenated and processed by a deeper, shared fusion encoder $\mathcal{E}_{\text{fuse}}$ to model high-level interactions (Li et al. 2021).

$$z_i = \mathcal{T}(\mathcal{E}_i(\Phi_i(x_i)))$$
$$g = \mathcal{T}(\mathcal{E}_{\text{fuse}}([z_1; z_2; \ldots; z_k]))$$

where Φ_i is a 1D convolution and $\mathcal{T}(\cdot)$ is an operator that selects the classification token's embedding.

Hierarchical Attention (One-to-Multi) Conversely, this architecture first models cross-modal interactions before refining modality-specific features. All inputs are projected by a linear layer Ψ_i and concatenated into a single sequence. This joint sequence is processed by a shared encoder $\mathcal{E}_{\text{shared}}$. The output sequence is then split into its original modality-specific segments, each of which is passed through a final dedicated encoder $\mathcal{E}_i$ (Lin et al. 2020).

$$h = \mathcal{E}_{\text{shared}}([\Psi_1(x_1); \ldots; \Psi_k(x_k)])$$
$$[h_1, \ldots, h_k] = \text{Split}(h)$$
$$g = [\mathcal{T}(\mathcal{E}_1(h_1)); \ldots; \mathcal{T}(\mathcal{E}_k(h_k))]$$

Cross-Attention Fusion (CAF) CAF facilitates direct, dense interaction between pairs of modalities. For a bimodal case (x_1, x_2), each modality's sequence is used to generate queries that attend to the keys and values of the other modality (Lu et al. 2019).

$$z_{1\leftarrow 2} = \text{MultiHead}(\Psi_Q(x_1), \Psi_K(x_2), \Psi_V(x_2))$$
$$z_{2\leftarrow 1} = \text{MultiHead}(\Psi_Q(x_2), \Psi_K(x_1), \Psi_V(x_1))$$
$$g = [\mathcal{T}(z_{1\leftarrow 2}); \mathcal{T}(z_{2\leftarrow 1})]$$

where Ψ_Q, Ψ_K, Ψ_V are modality-specific linear projection layers.

Cross-Attention Concatenation Fusion (CACF) CACF extends CAF by incorporating an additional global reasoning step (Zhan et al. 2021; Tsai et al. 2019). The cross-attended representations $(z_{1\leftarrow 2}, z_{2\leftarrow 1})$ are concatenated with initial linear projections of the original inputs $(x_i' = \Psi_i(x_i))$. This combined sequence is then processed by a final global Transformer encoder $\mathcal{E}_{\text{global}}$.

$$f = [x_1'; z_{1\leftarrow 2}; x_2'; z_{2\leftarrow 1}]$$
$$g = \mathcal{T}(\mathcal{E}_{\text{global}}(f))$$

Hybrid Logit Fusion Methods

We also implement two representative methods that operate directly on the output logits $\{\ell_i\}_{i=1}^k$ from modality-specific classifiers. Other methods can also be easily and quickly incorporated into our proposed benchmark.

Logit Summation (LS) This is the most direct parameter-free method for logit fusion. It operates under the assumption that each modality contributes equally to the final prediction. The logit vectors from all modality-specific classifiers are simply summed to produce the final fused logits: $\ell_{\text{fused}} = \sum_{i=1}^k \ell_i$.

Dataset	Concat	TF	Concat Early	LFT	EFT	Multi-to-One	One-to-Multi	CAF	CACF	LS	TMC
Houston2013	75.68	76.53	75.95	60.54	24.68	78.54	79.22	74.92	83.32	79.43	79.12
Houston2018	70.99	74.46	70.07	63.41	35.00	76.52	70.89	68.71	77.66	80.52	77.33
MUUFL Gulfport	84.21	83.90	86.26	71.77	46.68	80.95	81.23	83.99	86.42	86.64	83.20
Trento	98.43	96.95	98.14	95.68	71.64	97.71	96.08	97.74	98.68	98.53	97.67
Berlin	68.25	70.22	72.53	61.72	60.74	73.75	71.22	76.31	77.42	78.61	77.67
Augsburg	89.24	85.86	89.50	82.88	57.29	85.86	86.80	87.92	89.05	89.49	87.21
ForestNet	45.18	45.63	45.68	47.19	44.33	45.78	45.58	45.03	45.93	46.08	45.68

Table 2: Performance on Remote Sensing Datasets.

Dataset	Concat	TF	Concat Early	LFT	EFT	Multi-to-One	One-to-Multi	CAF	CACF	LS	TMC
TCGA-BRCA	78.17	77.18	77.18	69.44	65.67	76.79	76.19	78.77	78.37	77.18	75.40
ROSMAP	70.95	75.71	70.00	42.86	69.52	69.52	68.10	66.67	71.90	71.43	66.67
SIIM-ISIC	97.86	97.77	97.86	97.83	97.85	97.85	97.83	97.86	97.83	97.83	97.85
Derm7pt	45.49	52.72	45.41	43.96	38.86	45.92	46.77	48.13	52.72	46.34	52.72
GAMMA	61.43	62.86	63.81	62.38	59.05	57.62	65.71	63.33	63.33	62.38	61.43
MIMIC-III	68.09	68.81	68.48	68.42	68.59	69.01	68.76	68.51	68.88	68.47	68.87
eICU	90.05	90.03	90.05	90.05	90.05	90.05	90.05	90.07	90.05	90.05	90.03
TCGA	51.91	-	53.55	60.66	51.37	60.11	53.01	56.28	61.75	54.10	62.84
MIMIC-CXR (macro-F1)	0.6861	0.8458	0.6788	0.1551	0.1829	0.5229	0.4031	0.6136	0.7523	0.7644	-

Table 3: Performance on Medical AI Datasets. The dash "-" indicates the method is not applicable to this dataset.

Evidential Fusion (TMC) This method, based on evidential deep learning, transforms logits into evidence parameters α_i for a Dirichlet distribution (Han et al. 2022). This allows for explicit uncertainty quantification. The evidence from each modality is then fused using Dempster's rule of combination.

$$\alpha_i = \text{Softplus}(\ell_i) + 1, \alpha_{\text{fused}} = \bigoplus_{i=1}^{k} \alpha_i$$

Here, $\bigoplus$ denotes the Dempster-Shafer combination operator. The final class probabilities are derived from the fused evidence vector α_{fused}.

4.2 Automated Hyper-Parameter Tuning with Optuna

MULTIBENCH++ radically simplifies hyper-parameter tuning by using **Optuna**, eliminating traditional, GPU-heavy grid searches. A single `objective(trial)` callback efficiently handles the entire process, including:

- Dynamic search-space definition
- Module re-instantiation
- Early-stopping pruning
- Best checkpoint storage

This automated approach drastically cuts tuning time and resources, significantly boosting efficiency and performance.

Search-Space Specification For every trial, Optuna independently samples

- learning rate $\log\mathcal{U}(10^{-5}, 10^{-3})$,
- weight decay $\log\mathcal{U}(10^{-6}, 10^{-2})$,
- optimizer type $\in \{\text{AdamW}, \text{RMSprop}, \text{Adam}\}$.

Thus, the joint space spans three orders of magnitude in learning dynamics and two architectural regimes, while remaining compact for efficient Bayesian optimisation.

End-to-End Workflow Algorithm 1 demonstrates the end-to-end workflow. After retrieving a dataset via the unchanged data loader, one may substitute any of the presented Transformer fusion modules or logits-level combiners; the Optuna wrapper then orchestrates the hyperparameter search and returns a trained model, which is subsequently evaluated under the standard protocol.

5 Experiment and Discussion

5.1 Setup

Using MULTIBENCH++, we load each of the expanded datasets and systematically evaluate the multimodal approaches in our MULTIBENCH++ Algorithms mentioned in Sec. 4: We maintain a consistent experimental setup, varying only the method while keeping all other factors constant, including the training loop and data preprocessing steps. This approach ensures that observed differences in performance can be directly attributed to the fusion method under evaluation.

We compare our method with several classic baseline approaches previously proposed, including Concat, TensorFusion (`TF`), ConcatEarly, LateFusionTransformer (`LFT`), EarlyFusionTransformer (`EFT`) (Liang et al. 2021).

5.2 Overall performance

The performance metrics across diverse datasets highlight the nuanced effectiveness of different fusion strategies. As shown in Tables 2 to 5, we have the following observations: (i) Our algorithms (`CAF, CACF, Logit Summation,`

Dataset	Concat	TF	Concat Early	LFT	EFT	Multi-to-One	One-to-Multi	CAF	CACF	LS	TMC
MELD	61.34	65.66	62.73	47.77	57.92	66.37	64.02	65.54	62.22	61.60	65.56
IEMOCAP	54.96	54.59	54.82	32.64	48.14	54.49	54.16	54.26	55.06	54.30	51.22
MAMI	70.00	66.47	66.13	67.87	66.63	68.97	64.50	65.40	67.23	67.80	69.73
Memotion	77.89	77.75	78.14	78.09	78.09	78.09	78.09	78.14	78.09	78.19	78.09
MUTE	67.87	67.31	66.59	65.63	68.19	66.03	66.99	65.87	67.07	67.95	68.67
MultiOFF	56.95	59.86	55.38	57.76	59.11	54.38	59.33	59.76	62.03	54.16	60.01
MET-Meme(C)	35.67	34.87	36.92	29.68	24.12	33.55	35.09	32.89	36.62	34.94	33.85
MET-Meme(E)	42.95	42.47	41.83	32.69	29.81	44.23	39.74	40.38	41.03	43.11	42.47
Twitter2015	76.05	76.37	75.76	63.39	68.05	74.12	70.40	75.31	75.70	75.86	64.45
Twitter1517	76.83	76.72	76.68	76.76	76.68	76.29	76.72	76.54	75.97	76.15	76.86
CHSIMS(MSE)	0.4835	0.7431	0.4790	0.4775	0.4804	0.4772	0.4836	0.4794	0.4824	0.4898	-
CHSIMS-v2(MSE)	0.3202	0.3566	0.3338	0.3214	0.3391	0.3351	0.3448	0.2801	0.3361	0.3360	-

Table 4: Performance on Affective Computing & Social Media Understanding Datasets. The dash "-" indicates the method is not applicable to this dataset.

Dataset	Concat	TF	Concat Early	LFT	EFT	Multi-to-One	One-to-Multi	CAF	CACF	LS	TMC
MIRFLICKR	62.81	62.33	62.42	50.25	38.02	58.44	59.95	59.13	62.51	62.45	62.00
CUB Image-Caption	77.90	76.26	78.04	2.89	1.81	71.90	69.19	21.09	73.95	79.48	77.73
SUN-RGBD	60.28	59.43	60.83	45.22	31.61	53.52	56.97	53.53	58.14	60.78	59.00
NYUDv2	59.02	61.47	58.82	47.96	31.70	59.58	63.20	60.55	64.02	66.87	66.16
UPMC-Food101	91.95	92.04	91.80	83.76	8.75	86.04	88.80	86.89	90.12	91.66	92.02
MVSA-Single	79.83	78.36	78.29	68.40	63.39	79.32	77.78	79.51	78.03	79.25	67.50
MNIST-SVHN	96.41	96.64	96.42	62.11	93.06	95.45	93.47	95.35	96.45	96.46	96.95
N-MNIST+N-TIDIGITS	94.99	94.28	95.26	80.99	30.45	93.52	94.34	94.06	95.26	94.88	94.23
E-MNIST+EEG	58.72	58.21	61.28	17.69	7.95	42.56	30.51	49.74	59.49	62.05	57.69

Table 5: Performance on Other Datasets.

`TMC`, `One-to-Multi`, `Multi-to-One`) yield the highest accuracy on 26 of 37 datasets, routinely beating plain concatenation. (ii) Early fusion like `Concat` and `TF` collapses on weakly-aligned modalities, yet shows no gain on saturated tasks (**SIIM-ISIC**, **eICU**). (iii) `CACF` tops six benchmarks, confirming its broad efficacy.

Not surprisingly, the marginal utility of advanced fusion is strictly positive when and only when cross-modal redundancy is low; otherwise, naive concatenation attains near-optimal performance once any single modality approaches the task ceiling. Full results are provided in the appendix.

5.3 Data Complexity as a Model Selector

An analysis of the performance metrics in the Tables 2 to 5 reveals that data complexity is a critical factor for model selection. Taking Table 2 as an example, on the low-complexity **Trento** dataset, a simple model like `Concat` (98.43) is highly effective and performs nearly as well as the top model, `CACF` (98.68), indicating that increased model complexity provides little benefit. Conversely, for a high-complexity dataset like **Berlin**, there's a vast performance gap; simple models fail (`Concat` at 68.25) while sophisticated models like `LS` (78.61) and `TMC` (77.67) are essential for achieving high accuracy. This proves that the optimal model choice is not universal; it is dictated by the dataset's inherent complexity, requiring simple models for simple data and advanced architectures for complex data.

6 Future Work and Conclusion

Multimodal fusion's future hinges on two challenges: datasets lack fine-grained alignment for real validity, and models remain fragmented and unscalable. The path forward requires creating datasets with deeper structural correspondences while developing unified, theoretically-grounded fusion frameworks. Ultimately, this evolution must extend to the evaluation process itself, shifting from simple tuning towards ethics-aware meta-learning where fairness and robustness are primary objectives.

In conclusion, we present MULTIBENCH++, a rigorously-curated, open-source benchmark that unites 30+ datasets across 15+ modalities and 20+ tasks across specialized domains. Coupled with auto-tuned Transformer and hybrid-logit baselines, it gives researchers a fair testbed to compare new fusion models, making results easier to reproduce and closer to real-world use.

7 Acknowledgements

This work was supported by the National Natural Science Foundation of China (624B2100, 62376193, 61925602, T2522008, 62272055, 82522048 and 62501406), New Cornerstone Science Foundation through the XPLORER PRIZE. The authors acknowledge the MultiBench (Liang et al. 2021) team for their valuable contribution in developing and maintaining the open-source benchmark suite used in this work.

References

Akiba, T.; Sano, S.; Yanase, T.; et al. 2019. Optuna: A next-generation hyperparameter optimization framework. In *KDD*, 2623–2631.

Anumula, J.; Neil, D.; Delbruck, T.; and Liu, S.-C. 2018. Feature representations for neuromorphic audio spike streams. *Frontiers in neuroscience*, 12: 23.

Atrey, P. K.; Hossain, M. A.; El Saddik, A.; et al. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6): 345–379.

Azam, M. A.; Khan, K. B.; Salahuddin, S.; et al. 2022. A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics. *Computers in biology and medicine*, 144: 105253.

Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2018. Multimodal machine learning: A survey and taxonomy. *TPAMI*, 41(2): 423–443.

Bennett, D. A.; Buchman, A. S.; Boyle, P. A.; et al. 2018. Religious orders study and rush memory and aging project. *Journal of Alzheimer's disease*, 64(s1): S161–S189.

Busso, C.; Bulut, M.; Lee, C.-C.; et al. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4): 335–359.

Caesar, H.; Bankiti, V.; Lang, A. H.; et al. 2020. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 11621–11631.

Chen, D.; Su, W.; Wu, P.; and Hua, B. 2023. Joint multimodal sentiment analysis based on information relevance. *Information Processing & Management*, 60(2): 103193.

Cohen, G.; Afshar, S.; Tapson, J.; and Van Schaik, A. 2017. EMNIST: Extending MNIST to handwritten letters. In *IJCNN*, 2921–2926. IEEE.

Debes, C.; Merentitis, A.; Heremans, R.; et al. 2014. Hyperspectral and LiDAR data fusion: Outcome of the 2013 GRSS data fusion contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(6): 2405–2418.

Fersini, E.; Gasparini, F.; Rizzi, G.; et al. 2022. SemEval-2022 Task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, 533–549.

Gader, P.; Zare, A.; Close, R.; Aitken, J.; and Tuell, G. 2013. MUUFL Gulfport hyperspectral and LiDAR airborne data set. *Univ. Florida, Gainesville, FL, USA, Tech. Rep. REP-2013-570*.

Goyal, Y.; Khot, T.; Summers-Stay, D.; et al. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 6904–6913.

Han, Z.; Zhang, C.; Fu, H.; and Zhou, J. T. 2022. Trusted multi-view classification with dynamic evidential fusion. *TPAMI*, 45(2): 2551–2566.

Hossain, E.; Sharif, O.; and Hoque, M. M. 2022. MUTE: A multimodal dataset for detecting hateful memes. In *Proceedings of the 2nd conference of the asia-pacific chapter of the association for computational linguistics and the 12th international joint conference on natural language processing: student research workshop*, 32–39.

Hu, J.; Liu, R.; Hong, D.; et al. 2023. MDAS: A new multimodal benchmark dataset for remote sensing. *Earth System Science Data*, 15(1): 113–131.

Huiskes, M. J.; and Lew, M. S. 2008. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, 39–43.

Irvin, J.; Sheng, H.; Ramachandran, N.; et al. 2020. Forestnet: Classifying drivers of deforestation in indonesia using deep learning on satellite imagery. *arXiv preprint arXiv:2011.05479*.

Johnson, A. E.; Pollard, T. J.; Berkowitz, S. J.; et al. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1): 317.

Johnson, A. E.; Pollard, T. J.; Shen, L.; et al. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1): 1–9.

Kawahara, J.; Daneshvar, S.; Argenziano, G.; et al. 2018. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE journal of biomedical and health informatics*, 23(2): 538–546.

Kong, J.; Cooper, L. A. D.; Wang, F.; et al. 2011. Integrative, Multi-modal Analysis of Glioblastoma Using TCGA Molecular Data, Pathology Images and Clinical Outcomes. *T-BME*, 58(12): 3469–3474.

Li, R.; Yang, S.; Ross, D. A.; and Kanazawa, A. 2021. Ai choreographer: Music conditioned 3d dance generation with aist++. In *ICCV*, 13401–13412.

Li, Y.; Li, Y.; Wang, X.; et al. 2024. Benchmarking Multimodal Retrieval Augmented Generation with Dynamic VQA Dataset and Self-adaptive Planning Agent. In *ICLR*.

Liang, P. P.; Lyu, Y.; Fan, X.; et al. 2021. Multibench: Multiscale benchmarks for multimodal representation learning. *NeurIPS*, 2021(DB1): 1.

Lin, J.; Yang, A.; Zhang, Y.; Liu, J.; Zhou, J.; and Yang, H. 2020. Interbert: Vision-and-language interaction for multimodal pretraining. *arXiv preprint arXiv:2003.13198*.

Lin, N.; Wang, S.; Li, Y.; et al. 2025. Resistive memory-based zero-shot liquid state machine for multimodal event data learning. *Nature Computational Science*, 5(1): 37–47.

Liu, Y.; Yuan, Z.; Mao, H.; et al. 2022. Make acoustic and visual cues matter: Ch-sims v2. 0 dataset and av-mixup consistent module. In *ICMI*, 247–258.

Lu, D.; Neves, L.; Carvalho, V.; Zhang, N.; and Ji, H. 2018. Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1990–1999.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS*, 32.

Nagrani, A.; Yang, S.; Arnab, A.; et al. 2021. Attention bottlenecks for multimodal fusion. *NeurIPS*, 34: 14200–14213.

Niu, T.; Zhu, S.; Pang, L.; and El Saddik, A. 2016. Sentiment analysis on multi-view social data. In *International conference on multimedia modeling*, 15–27. Springer.

Okujeni, A.; van der Linden, S.; and Hostert, P. 2016. Berlin-urban-gradient dataset 2009-an enmap preparatory flight campaign.

Orchard, G.; Jayawant, A.; Cohen, G. K.; and Thakor, N. 2015. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in neuroscience*, 9: 437.

Pollard, T. J.; Johnson, A. E.; Raffa, J. D.; et al. 2018. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific data*, 5(1): 1–13.

Poria, S.; Hazarika, D.; Majumder, N.; et al. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *ACL*, 527–536.

Ramachandram, D.; and Taylor, G. W. 2017. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, 34(6): 96–108.

Rotemberg, V.; Kurtansky, N.; Betz-Stablein, B.; et al. 2021. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific data*, 8(1): 34.

Sharma, C.; Bhageria, D.; Scott, W.; et al. 2020. SemEval-2020 Task 8: Memotion Analysis-the Visuo-Lingual Metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 759–773.

Shi, Y.; Paige, B.; Torr, P.; et al. 2019. Variational mixture-of-experts autoencoders for multi-modal deep generative models. *NeurIPS*, 32.

Silberman, N.; Hoiem, D.; Kohli, P.; and Fergus, R. 2012. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, 746–760. Springer.

Soleymani, M.; Garcia, D.; Jou, B.; et al. 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65: 3–14.

Song, S.; Lichtenberg, S. P.; and Xiao, J. 2015. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, 567–576.

Suryawanshi, S.; Chakravarthi, B. R.; Arcan, M.; et al. 2020. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, 32–41.

Tsai, Y.-H. H.; Bai, S.; Liang, P. P.; et al. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, 6558.

University of Trento. 2022. Theses of the University of Trento. [Data set]. Original work published 2020.

Wang, X.; Kumar, D.; Thome, N.; Cord, M.; and Precioso, F. 2015. Recipe recognition with large multimodal food dataset. In *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 1–6. IEEE.

Wang, Y.; Chen, X.; Cao, L.; et al. 2022. Multimodal Token Fusion for Vision Transformers. In *CVPR*.

Wei, X.; Zhang, T.; Li, Y.; et al. 2020. Multi-Modality Cross Attention Network for Image and Sentence Matching. In *CVPR*.

Weinstein, J. N.; Collisson, E. A.; Mills, G. B.; et al. 2013. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10): 1113–1120.

Willett, F. R.; Avansino, D. T.; Hochberg, L. R.; et al. 2021. High-performance brain-to-text communication via handwriting. *Nature*, 593(7858): 249–254.

Wu, J.; Fang, H.; Li, F.; et al. 2023. Gamma challenge: glaucoma grading from multi-modality images. *Medical Image Analysis*, 90: 102938.

Xu, B.; Li, T.; Zheng, J.; Naseriparsa, M.; et al. 2022. Metmeme: A multimodal meme dataset rich in metaphors. In *SIGIR*, 2887–2899.

Xu, P.; Zhu, X.; and Clifton, D. A. 2023. Multimodal learning with transformers: A survey. *TPAMI*, 45(10): 12113–12132.

Xu, Y.; Du, B.; Zhang, F.; and Zhang, L. 2018. Hyperspectral image classification via a random patches network. *ISPRS journal of photogrammetry and remote sensing*, 142: 344–357.

Yu, W.; Xu, H.; Meng, F.; et al. 2020. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *ACL*, 3718–3727.

Zadeh, A.; Zellers, R.; Pincus, E.; and Morency, L.-P. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

Zadeh, A. B.; Liang, P. P.; Poria, S.; et al. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2236–2246.

Zhan, X.; Wu, Y.; Dong, X.; et al. 2021. Product1m: Towards weakly supervised instance-level product retrieval via cross-modal pretraining. In *ICCV*, 11782–11791.

Zhang, Q.; Fu, J.; Liu, X.; and Huang, X. 2018. Adaptive co-attention network for named entity recognition in tweets. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Zhang, Q.; Wei, Y.; Han, Z.; et al. 2024. Multimodal fusion on low-quality data: A comprehensive survey. *arXiv preprint arXiv:2404.18947*.

Zhang, Q.; Wu, H.; Zhang, C.; et al. 2023. Provable dynamic fusion for low-quality multimodal data. In *ICML*, 41753–41769. PMLR.

Zhu, J.; Zhou, Y.; Qian, S.; et al. 2025. Mosaic of modalities: A comprehensive benchmark for multimodal graph learning. In *CVPR*, 14215–14224.