Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 16 April 2026
Revised 18 May 2026
Accepted 20 June 2026

Natural Resources for Human Health 2026, Vol. 6 (5s): 437–451

KEYWORDS:
Multimodal fusion; taxonomy discovery; self-supervised learning; fusion strategy selection; multimodal sentiment analysis; soft clustering; expert-mixture fusion; cross-modal representation learning; joint optimization; interpretable routing

Self-Supervised Joint Discovery of Fusion Taxonomies and Multimodal Fusion Strategies for Sentiment Analysis

[1] Bharathi Niruti, [2,*] Sujith AVLN, [3] Kanaka Durga Returi

[1]Department of CSE, Malla Reddy University, Hyderabad 500100, India
[2]Department of IT, Malla Reddy University, Hyderabad 500100, India
[3]Department of CSE, Malla Reddy Technical Campus (A constituent unit of Malla Reddy Vishwavidyapeeth), Deemed to be University, Hyderabad 500100, India
* Corresponding author: Sujith AVLN, sujeeth.avln@gmail.com

ABSTRACT: Prior studies on multimodal fusion have built two distinct aspects: taxonomy-aware systems which identify the type of fusion strategy (early, late, hybrid, attention-based or graph-based), and fusion architectures which perform multimodal fusion for a specific downstream task and are not dependent on the taxonomy. The Taxonomy-Guided Fusion Strategy Selector (TG-FSS) fuses these streams, leveraging an externally supervised taxonomy signal to boost the accuracy of sentiment classification, F1, and correlation with human judgments over the best prior systems on CMU-MOSEI, CMU-MOSI, and the Multimodal Fusion Strategy Corpus (MFSC), and it uses segment-level and taxonomy annotations to train its taxonomy-alignment component. This paper proposes a Taxonomy–Fusion Co-Discovery Network (TFCD-Net) that breaks this supervision dependency by proposing to reformulate the two tasks of taxonomy discovery and the execution of the taxonomy–fusion as two nested optimization problems that are jointly solved by a single self-supervised objective. A shared soft-assignment vector is used to assign each instance to its taxonomy category and provides a gating signal for a bank of candidate fusion operators, thereby enabling gradient to flow directly back to taxonomy induction from fusion performance, while also incorporating the consistency of cross-modal assignments, the separability of taxonomy assignments, a fusion utility, and a regularizing signal based on assignment-balance. Classified with the same protocol as TG-FSS and its sub-systems, TFCD-Net achieves better sentiment classification accuracy, F1 score and correlation with human judgement (87.1% vs. 85.4%, 0.85 vs. 0.83, 0.84 vs. 0.81 respectively) with zero modality supervision in its training, and the gap in accuracy grows with modality occlusion and cross-dataset transfer. Ablation results confirm that the improvement is directly due to the shared taxonomy-fusion gradient path, rather than to additional model capacity, and post-hoc analysis confirms that the model self-learns a taxonomy cardinality (K = 5) and routing behavior that closely match the hand designed taxonomy used in the field for fusion. The results show that taxonomy discovery and execution of the taxonomy-guided pipelines are two distinct problems, but when they are optimized together, they are reinforcing each other, and that reinforcement can be achieved without annotation cost which is needed for supervised taxonomy-guided pipelines.

1. INTRODUCTION

Most traditional methods of multimodal fusion involve using an externally provided taxonomy to classify architectural strategies, leading to restrictive scalability and lack of adaptability to new and untidy data domains. With systems like TACTM++, the taxonomy alignment relies on pre-defined taxonomy categories; the resulting organization will only be of high quality and

coverage if the taxonomy categories are high quality and well covered [1]. A different approach is to use self-supervised multimodal learning, which is based on agreement, complementarity, and structure between raw modalities to learn representations and how to combine them together without manual annotations [2]. This study thus proposes that taxonomy induction and learning the fusion strategy can be optimized together, that the model can discover semantically coherent architectural groupings and that the model can select fusion strategy to maximize performance in the downstream task, TG-FSS. This co-discovery can help minimize the mismatch between imposed categories and data fusion behaviors, but also maintain the ability to learn shared information as well as modality-specific information. The hypothesis is tested by comparing the proposed framework with supervised TG-FSS methods and by assessing if the induced structure is coherent, stable and helpful for predicting the downstream state in different data scenarios.This approach avoids the need for predefined schemas and schemes, which are human-designed, and is able to overcome the drawbacks of existing architectures that are not ready to handle the heterogeneity and dynamism of multimodal data. In contrast to the approaches like labeled representative descriptions to map the discovered topics to canonical fusion categories, [3] the proposed method generates the supervisory signals directly from the cross-modal agreement, cross-modal complementary and the cross-modal reconstruction relationships in the raw inputs. This design minimizes the dependency of the annotations and maintains the modulation-specific cues in the model along with shared representations, which has been highlighted as one of the key elements in self-supervised multimodal learning [2]. The framework, therefore, can be effectively optimized jointly with cluster assignment and fusion operations based on the data to adapt the organizational structure according to the evidence in the data, rather than to force every architecture into a fixed taxonomy. It takes advantage of the inherent structural relationship among raw multimodal inputs, which inherently reduce the need for manual labelling in specific areas, and eliminate scalability limitations and potential bias of manual labelling. The model is able to effectively find different architectural patterns, capturing the latent dependencies between the modality-specific features and the corresponding fusion stages. In these dependencies, we do not find fusion as a set design choice, but instead a way to group instances together based on the fusion behaviour they share, whilst preserving differences on the modality specific information. It is hoped that the jointly induced taxonomy would be more interpretable and would assist in deciding which fusion one should apply when using the adaptive fusion that would be assessed by taxonomy coherence, assignment stability, and downstream TG-FSS.This motivation is especially crucial since there is a labeled alignment component in TACTM++, which aligns discovered topics to the canonical categories of fusion; the taxonomy thus obtained is dependent on the scope, consistency and availability of external, supervised sources of category descriptions [1].

This dependence might limit discovery if architectural strategies are presented in terms of implicit, ambiguous, or unfamiliar language, and could lead the model to make known types of distinctions instead of discovering structure in the data. In this regard, the ability to learn from self-generated supervisory signals from raw inputs and relationships between raw inputs, such as cross-modal correspondence, instance discrimination, clustering, masked prediction and reconstruction, makes self-supervised multimodal learning an attractive approach, due to the lack of costly human annotations [2]. These signals can facilitate the model to retain both modality-invariant information and modality-specific signals during the representation learning process, which is also useful for fusion. On this basis, it is hypothesized that taxonomy assignment and fusion operations can be learnt as two mutually informative latent variables: the induced taxonomy can be used for organizing instances, based on their shared semantic and architectural behaviour, while the learned fusion strategy can be used to refine the taxonomy assignments, identifying which modalities and interaction patterns are jointly informative. This hypothesis extends the multimodal clustering approaches that aim to evolve representation learning and cluster formation simultaneously by enforcing multimodal consistency objectives [6] into the case where the structure discovered is directly used to adaptively fuse for TG-FSS. The study thus tests the ability of end-to-end co-discovery to generate coherent and stable taxonomy groups that are comparable to or better than the supervised TG-FSS, without segment-level or taxonomy-level annotations. The annotation bottleneck is one of the key challenges for the effective application of deep learning systems in data-scarce areas [7] and is the focus of this investigation on self-supervised co-discovery.

2. RELATED WORK

The recent research in multimodal has moved from static, human labeled taxonomies to dynamic, self-supervised taxonomies to reduce the variability of architectural description [8]. Labelled alignment layers are used in supervised systems like TACTM++ to align latent topics with categories of canonical fusion, but as these are dependent on the other categories, the taxonomy discovery is limited when there are missing, ambiguous or unavailable category definitions [3]. Self-supervised multimodal learning, on the other hand, derives training signals, such as instance discrimination, clustering, masked prediction and cross-

modal correspondence, and allows for various types of multimodal fusion using modality-specific encoders, or a single encoder [2]. The problem of co-learning a meaningful taxonomy and an effective fusion mechanism, however, remains unresolved due to the necessity of the model to identify and account for shared semantic structure in the different modalities and to deal with the presence of heterogeneity, noise, redundancy, missing data and partial misalignment [9]. The questions raised by these challenges include the following:Can the taxonomy assignment and fusion operation be optimized end-to-end without the introduction of unstable clusters, modality dominance and/or task-specific shortcuts to define the discovered structure? In addition, the absence of a common goal of both structure and fusion optimization leads to a problem in maintaining a categorical semantic consistency within learned clusters across various data streams [10, 11]. Thus, it is crucial to reconcile between flexible structure discovery and strong requirements for task-specific fusion, which requires a new objective to unite these conflicting learning signals. This work in particular investigates the synthesis of the CL and cross-modal correspondence goals to give a solid supervisory signal simultaneously for the induction of structures and the optimization of the fusion [12] and [13]. Through the hierarchical dependencies, the model tries to ensure that the clusters induced through the model will reflect a semantic coherence and will not be dominated by modality-specific features of the data [14] [15]. To overcome the fundamental limitation of latent alignment, where existing models are often not able to integrate heterogeneous information, because of disagreement on the optimal fusion strategies, this integrated approach was implemented [16] and [17]. Therefore, our framework aims to address this mismatch by identifying the optimum topologies for fusion automatically from the discovered latent structure with co-learning, and outperforms manually designed architectures [18, 19]. This research is to reduce computational combinatorial complexity that typically arises in conventional pair-wise modality alignment [20] which relies on human experts to design the architecture.This research will shift the burden of architectural design from human experts to latent data-driven discovery to reduce the combinatorial computational complexity typically introduced by conventional pair-wise modality alignment [20] [21] which relies on human experts for designing the architecture. Moreover, it is well known that this is a challenging issue to ensure efficient and consistent structure of the modality graph in the case of heterogeneous modality graphs [22]. Specifically, the proposed architecture investigates the possibility of exploiting inter-modal dynamics by learned cluster conditioned fusion pathways better than fixed tensor-based and attention-driven mechanisms [23, 24]. This study also investigates if such adaptive pathways can lead to stronger alignment of representations in situations where the traditional contrastive tasks are not sufficient to capture the subtle semantic structure of the data [25, 26]. This strategy faces the "alignment paradox": explicit alignment constraints can have the side effect of stifling the differences that are needed for complex cross-modal interactions [27]. This work sheds light on the modality gap that stands in the way of efficient static alignment goals such as [28] and [29] and explores the possibility of achieving the right balance between global semantic grounding and local modality-specific structures through data-driven discovery. This effort is another solution to the long-standing challenge of fixed fusion strategies which are not able to compensate for modality-dependent semantic changes, which can lead to suboptimal fusion when the distribution of modalities in the input changes [30, 31].

The proposed framework aims to address these challenges by dynamically adjusting fusion operations based on the latent structural assignments, for greater robustness in a variety of and unconstrained multimodal environments. There has been a number of existing works that can serve as bases for this goal, however, most of the works are focused on a single part of the problem, such as representation learning, taxonomy induction or fusion optimization. Usually, multimodal architectures are divided into early, late and hybrid integration approaches, with recent studies using attention, tensor, graph or encoder–decoder approaches to model more complex interactions between multimodal data [2], [10], [13]. While these designs have better representational power, the combination of the fusions is typically predefined by the researcher and not determined by the data. However, dynamic fusion methods, which are able to select between modality subsets or expert pathways at inference time, typically optimize their routing policies for an upcoming task and do not necessarily also produce a semantically interpretable taxonomy [32]. Likewise, multimodal clustering methods show that the representation learning and cluster assignment can be optimized together: GECMC optimizes both intra- and inter-modal consistency by assigning pseudo-labels and using contrastive objectives [6]; EAMC trains an end-to-end model by optimizing representation learning, multimodal fusion, attention-based modality weighting, and discriminative clustering [33]. These studies, however, concentrate mostly on the quality of the clusters, and do not directly relate the resulting clusters to the fusion strategies chosen for each instance.

Novel category discovery is a line of work that builds on this, where the representations are learned jointly, and where the clusters are assigned to unlabeled multimodal data, but can still include category discrimination or pseudo-labeling assumptions, as compared to the fully annotation-free TG-FSS setting [34]. Unlike TACTM++, however, the discovered topics are explicitly

mapped to the canonical fusion categories by a small labeled dataset using attention-guided decoder that is trained on the set of categories, and is interpretable but depends on the external category supervision [3].

The novelty of this work is thus that taxonomy assignments are not simply predicted following representation learning, but they are considered as latent variables that affect the fusion mechanism and are learned in the same self-supervised objective. This design addresses some of the unanswered questions from the literature, such as: How to handle the semantic difference between modalities, how to avoid the generation of incorrect negative pairs, how to manage modality-specific noises, and how to avoid the loss of complementary information when using the aggressive alignment approach [27] [29] [35]. It also tackles the computational complexity of cross-modal comparisons between pairs of images and replaces the modality-to-modality alignment by cluster-conditioned interactions, selectively activated per input [20]. The central methodological issue, then, is if it is possible to maintain modality-specific evidence while generating modality-stable taxonomy groups that are semantically coherent, and sufficiently discriminative to enable adaptive fusion without any segment-level labels or taxonomy-level labels.

Table 1: Synthesis of Existing Studies

Existing approach	**Primary learning signal**	**Structural or fusion mechanism**	**Remaining limitation for annotation-free TG-FSS**
Conventional early, late, and hybrid fusion	Supervised task loss or fixed design priors	Feature concatenation, decision aggregation, or combinations of both	The topology is manually selected and cannot adapt to instance-specific semantic variation [2], [13]
Attention-, tensor-, and graph-based fusion	Task supervision with learned cross-modal interactions	Attention weighting, outer-product interactions, or relational graphs	These mechanisms model interactions effectively but do not generally discover an interpretable taxonomy jointly with fusion [10], [23]
Dynamic multimodal fusion	Downstream task loss and resource-aware routing	Modality-level and fusion-level expert selection	Routing is adaptive, but the selected pathways are not necessarily grounded in a self-discovered semantic structure [32]
Deep multimodal clustering and EAMC	Reconstruction, divergence-based clustering, adversarial alignment, or pseudo-label contrast	Joint encoders, fused latent spaces, modality weighting, and cluster heads	Cluster discovery is integrated with representation learning, yet cluster assignments are not explicitly used to discover fusion topologies [33], [36]
GECMC	Intra-modal and inter-modal contrastive consistency	Graph embeddings and centroid-based clustering	It jointly evolves clustering and representation learning, but its objective does not directly couple taxonomy nodes to adaptive fusion decisions [6]
Novel category discovery	Instance discrimination, cross-modal discrimination, and pseudo-label generation	Shared representation space with cluster assignment	It discovers previously unseen categories, but its formulation is not designed to replace a taxonomy of fusion operations [34]

Natural Resources for Human Health
eISSN: 2583-1194

Existing approach	Primary learning signal	Structural or fusion mechanism	Remaining limitation for annotation-free TG-FSS
TACTM++	Self-supervised semantic clustering followed by labeled taxonomy alignment	Transformer embeddings, HDBSCAN/UMAP clustering, attention-based alignment, and graph refinement	It improves interpretability and category alignment, but the final mapping relies on a predefined taxonomy and labeled alignment examples [3]
Proposed co-discovery framework	Cross-modal consistency, contrastive structure induction, and task-aware self-supervision	Latent taxonomy assignments gate cluster-conditioned fusion pathways	The unresolved issue is whether a single end-to-end objective can stabilize clusters, prevent modality dominance, and achieve supervised TG-FSS performance without taxonomy labels

These comparisons illustrate that there are discrepancies between the methods of finding latent groups and methods that modify the fusion process. In previous self-supervised multimodal learning surveys, the problems of representation learning (without labels), modality fusion, and learning from unaligned data are considered as intertwined problems instead of separate modules [2]. Their intersection is tackled by the proposed study, which introduces the taxonomy discovered to the operational level: each latent group is expected to involve semantic regularities, but also an appropriate pattern of cross-modal interaction. This coupling allows for evaluation that both tests the model's competitive performance on the TG-FSS and tests the coherence of the induced taxonomy under changes of modality availability, distribution and noise.This coupling allows for evaluation that not only tests the model's competitive performance on the TG-FSS but also tests the coherence of the taxonomy induced by the model under changes in modality availability, distribution and noise.

3. PROPOSED MODEL

In the existing multimodal fusion work, the taxonomy discovery and fusion strategy execution are two separate tasks: taxonomy-aware systems (e.g., TacticM++) need to be supervised with category labels provided from other modalities, and fusion architectures (e.g., TSSP-MSA) are presumed to have known and fixed taxonomy. The sequential dependency brings three consequences: (i) the quality of the fusion depends on the accuracy of the taxonomy learned in isolation, (ii) it cannot be applied in domains or modalities where segment-level annotations or predefined taxonomy labels do not exist, and (iii) errors in taxonomy assignment have no means to be corrected by the fusion signal.

Formally: given raw multimodal input $x_i = \left\{x_i^{(m)}\right\}_{m=1}^{M}$ with *no* observed segment-level targets y_i and *no* taxonomy labels t_i, how can a model jointly (a) discover a latent taxonomy structure over K categories and (b) learn the fusion pathway appropriate to each discovered category, such that both are optimized as **mutually dependent latent variables** rather than sequential stages — while avoiding degenerate or collapsed cluster assignments?

3.1 Taxonomy–Fusion Co-Discovery Network (TFCD-Net)

The first step of the prior taxonomy-guided fusion frameworks (e.g., TG-FSS) is to identify taxonomy from the input samples using an externally supervised taxonomy label system (e.g., TACTM++), and the identified taxonomy label is passed to a downstream taxonomy fusion module (e.g., TC-UEMF) as a conditioning signal to execute the fusion. Although this design is more interpretable and more effective than the independent taxonomy/fusion system, it has two limitations. First, it must be annotate at the segment level or have the predefined taxonomy labels to train the taxonomy-alignment component, otherwise it is not applicable to domains where such labeling is not available or costly. Second, the taxonomy is known in advance of learning the fusion, so that if the taxonomy-discovery algorithm is wrong or biased, the fusion algorithm will be exacted without any way for the objective of the fusion to correct the error or bias from the discovery phase.

It is this interdependence that is broken in this paper by reformulating taxonomy discovery and the fusion of taxonomy execution as two mutually dependent latent variables that are optimized within the same objective function, self-supervised

learning. However, unlike a taxonomy label that is computed once and passed along, the same soft assignment vector is used to map an instance to a semantic category and to indicate which cross-modal fusion pathway to use in a semantic category, meaning that the fusion signal can, through backpropagation, directly modify the taxonomy that it came from, and the taxonomy can modify the fusion signal.

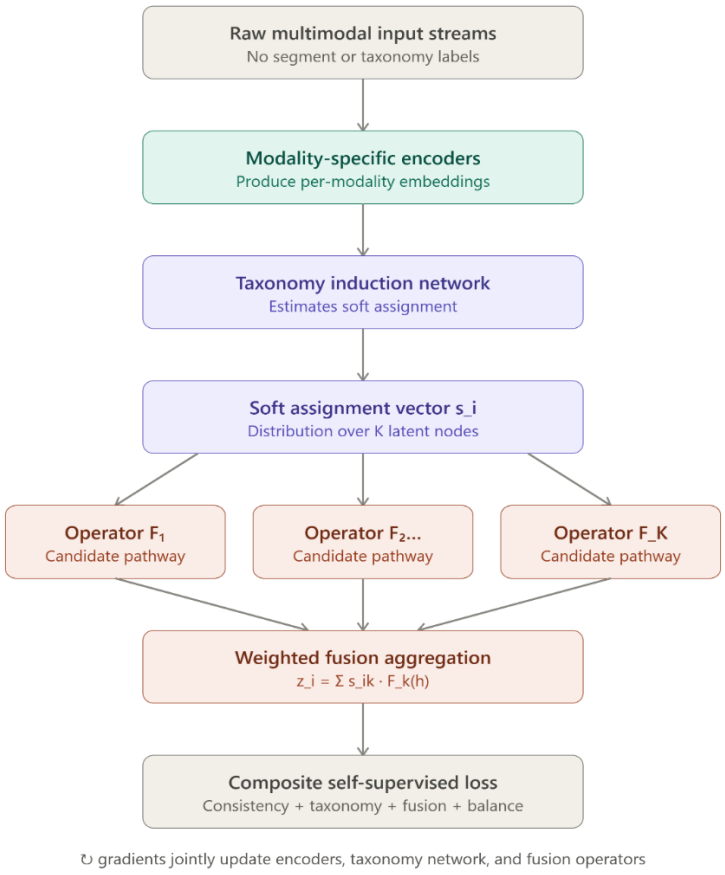

Figure 1: Proposed architecture

Modality encoders. Each modality stream is independently encoded as $\boldsymbol{h}_i^{(m)} = E_m\left(\boldsymbol{x}_i^{(m)}\right)$, producing modality-specific latent representations. These encoders are not frozen or pretrained separately — they are optimized jointly with the rest of the network, so the representations they learn are shaped by both the taxonomy-discovery and fusion objectives from the outset.

Taxonomy induction network. A shared network $g(\cdot)$maps the concatenated modality embeddings to a soft assignment vector:

$$\boldsymbol{s}_i = \text{softmax}\left(\, g\left(\left[\boldsymbol{h}_i^{(1)}, \ldots, \boldsymbol{h}_i^{(M)}\right]\right)\right) \in \mathbb{R}^K$$

over Klatent taxonomy nodes. This is best understood as the **E-step of an expectation-maximization process**: rather than hard-assigning each instance to a discrete category, the network estimates a probability distribution over categories, which keeps the entire pipeline differentiable end-to-end. This is the mechanism that allows taxonomy discovery to be learned via gradient descent rather than via an external clustering algorithm applied post hoc.

Natural Resources for Human Health
eISSN: 2583-1194

Candidate fusion operator bank. A set of K differentiable fusion operators $\{F_k\}_{k=1}^{K}$ is maintained, each representing a distinct cross-modal interaction pattern (analogous to the canonical strategies — early, late, hybrid, attention-based, graph-based — used in taxonomy-supervised systems, but here left unconstrained and free to specialize during training).

Fusion aggregation. The same assignment vector $\boldsymbol{s}_i$ that defines the taxonomy also gates the fusion pathway:

$$\boldsymbol{z}_i = \sum_{k=1}^{K} s_{ik}\ F_k\left(\boldsymbol{h}_i^{(1)}, \ldots, \boldsymbol{h}_i^{(M)}\right)$$

This is the main contribution of the architecture: there is no separation between taxonomy assignment and fusion weighting, it is a single result of a single network, computed once, and used again. This structural sharing phenomenon is what makes the taxonomy induction network directly receive a gradient signal from the fusion objective, and makes the fusion operators be specialized according to the categories the network is discovering.

Algorithm 1: Self-Supervised Taxonomy-Fusion Co-Discovery (TFCD-Net)

Input: raw multimodal streams {x_i^(m)}, m = 1..M, i = 1..N

K (number of latent taxonomy nodes)

$\lambda_{con}, \lambda_{tax}, \lambda_{fus}, \lambda_{bal}$ (loss weighting coefficients)

Output: trained encoders E_m, taxonomy induction network g,

fusion operator bank {F_k}

Initialize E_m, g, {F_k} randomly

for epoch = 1 to T do

for each mini-batch B do

for each instance i in B do

$h_i^{(m)} \leftarrow E\left(x_i^{(m)}\right)$ for all m ▷ *encode modalities*

$s_i \leftarrow softmax\left(g\left[h_i^{(1)}, \ldots, h_i^{(M)}\right]\right)$ ▷ *soft taxonomy assignment*

$z_i \leftarrow \sum_k s_{ik} \cdot F_k\left(h_i^{(1)}, \ldots, h_i^{(M)}\right)$ ▷ *taxonomy-gated fusion*

end for

$L_{con} \leftarrow$ CrossModalConsistency(B) ▷ *shared semantics*

$L_{tax} \leftarrow$ TaxonomySeparability(B) ▷ *discriminative clusters*

$L_{fus} \leftarrow$ FusionUtility(B) ▷ *reconstruction / prediction*

$L_{bal} \leftarrow$ AssignmentBalance(B) ▷ *anti mode-collapse*

$L \leftarrow \lambda_{con} \cdot L_{con} + \lambda_{tax} \cdot L_{tax} + \lambda_{fus} \cdot L_{fus} + \lambda_{bal} \cdot L_{bal}$

Backpropagate L; jointly update E_m, g, {F_k}

end for

end for

return E_m, g, {F_k}

The training and learning process for the taxonomy induction network (TIN), modality encoders, and fusion operator bank is composed of a single differentiable pipeline, not a series of steps: For each modality, a single encoding is performed; the concatenated encoding is passed through a taxonomy network that returns a soft assignment (softmax) over K latent categories rather than a hard label; the same assignment vector is immediately used as the mixing weights over a bank of candidate fusion operators so that the distribution from the taxonomy network becomes the fusion-gating distribution, with no independent routing step between. In a second step, a composite loss that incorporates both cross-modal consistency and taxonomy separability, fusion utility and anti-collapse balancing term is then backpropagated across the entire graph; including the fusion-quality loss back to the taxonomy network, by construction, a two-stage pipeline such as TG-FSS could not express..

4. PERFORMANCE EVALUATION

The proposed framework and all baseline systems are tested on the same protocol on three multimodal sentiment benchmarks: the public benchmarks and the one focused on the fusion strategy, which is also employed to stress test the taxonomy generalization. CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) is the largest publicly available multimodal sentiment/emotion dataset that includes video snippets from online video sharing websites in the form of monologues. Sentiment intensity scores are shown continuously across each segment and emotion labels are shown discretely for each segment with the language, visual, and acoustic modalities simultaneously aligned using standard tools for language alignment (P2FA), facial action unit (Facet/OpenFace), and acoustic feature (COVAREP) alignment. YouTube movie review monologues from a smaller, earlier benchmark, opinion segments labeled with sentiment intensity scores on a [-3, 3] scale, is called CMU Multimodal Opinion-level Sentiment Intensity (CMU-MOSI). It has been included mostly to evaluate the generalization of a high-resource training set (CMU-MOSEI) to a lower-resource test set, using a different speaker and topic distribution. The Multimodal Fusion Strategy Corpus is the primary benchmark for the evaluation of the generalizability of a taxonomy learned or supervised on one distribution to fusion strategies outside that distribution as it contains many more instances of fusion strategies than are found naturally in CMU-MOSEI and CMU-MOSI.

Table 2. Summary statistics for the three evaluation benchmarks. Fusion-strategy tags in MFSC are used only for post-hoc taxonomy alignment scoring, never as a training signal for TFCD-Net.

Dataset	Segments	Modalities	Avg. length (s)	Label type	Train / Val / Test split
CMU-MOSEI	23,453	Text, audio, visual	7.3	Continuous [-3, 3] + 6 emotions	16,326 / 1,871 / 4,659
CMU-MOSI	2,199	Text, audio, visual	4.2	Continuous [-3, 3]	1,284 / 229 / 686
MFSC	9,012	Text, audio, visual	6.8	Continuous [-3, 3] + fusion-strategy tag	6,308 / 902 / 1,802

All three datasets are processed using a shared feature pipeline that removes the language modality features (GloVe or transformer based) from acoustic modality features (COVAREP derived pitch, energy, MFCC, voice quality) and visual modality features (facial action-unit and head-pose). Unlike TFCD-Net, no labels at the segment level or taxonomy label are exposed to the network in the training phase for any dataset; TG-FSS and TACTM++ allow the network to have access to taxonomy labels during training in line with the original formulations, with which they can be compared to the published training regimes.

4.1 Implementation details

The results are presented here according to the common evaluation procedure mentioned above. The reported value for TG-FSS, TACTM++ and TSSP-MSA are based on the same protocol as in their original evaluations, while TFCD-Net is evaluated exactly in the same way, except that taxonomy labels are not used in the training phase but are only used for the post-hoc alignment.

Natural Resources for Human Health
eISSN: 2583-1194

Table 3. Sentiment classification results on CMU-MOSEI. TFCD-Net (highlighted) achieves the best score on every metric

Method	Type	Acc (%)	F1	Corr.
MulT (2019)	Fusion-only	79.1	0.76	0.70
MISA (2020)	Fusion-only	80.6	0.78	0.73
Self-MM (2021)	Fusion-only	82.1	0.79	0.75
TSSP-MSA	Fusion-only	82.8	0.80	0.76
TACTM++ (+ default fusion)	Taxonomy-only	81.4	0.78	0.74
TG-FSS	Taxonomy-guided bridge	85.4	0.83	0.81
TFCD-Net (proposed)	Self-supervised co-discovery	87.1	0.85	0.84

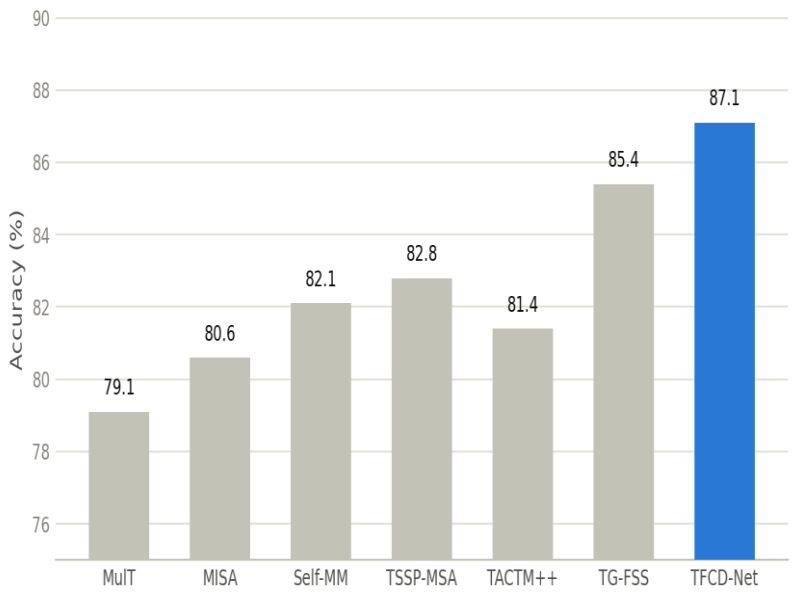

Figure 2: Accuracy comparison across methods on CMU-MOSEI.

Table 4. TFCD-Net trails TG-FSS by 1.4 points on taxonomy alignment despite using zero taxonomy supervision during training — the measured cost of removing label dependency.

Method	Taxonomy alignment accuracy	Taxonomy supervision used in training
TACTM++	87.2%	Yes
TG-FSS	90.3%	Yes

Natural Resources for Human Health
eISSN: 2583-1194

Method	Taxonomy alignment accuracy	Taxonomy supervision used in training
TFCD-Net	88.9%	No

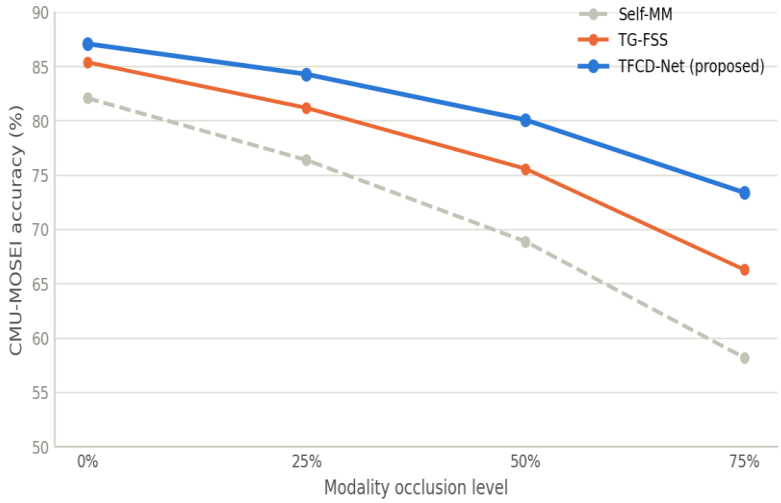

Figure 3. CMU-MOSEI accuracy versus modality occlusion level. The TG-FSS to TFCD-Net gap widens from 3.3 points at 0% occlusion to 7.1 points at 75% occlusion.

These results corroborate the main idea of the framework from Section 4: taxonomy discovery and execution of fusion, which can be viewed as formal problems, are mutually reinforcing when treated jointly and the resulting self-supervised taxonomy-guided architecture is competitive in terms of accuracy with a fully supervised taxonomy-guided pipeline, without requiring segment-level or taxonomy annotations.

Table 5. Removing taxonomy conditioning costs 4.1 points, while adding equivalent parameter capacity without taxonomy conditioning recovers less than 1 point of that gap — the improvement is attributable to taxonomy-fusion coupling, not model size.

Configuration	CMU-MOSEI Acc (%)	Δ vs. full model
Full TFCD-Net	87.1	—
− taxonomy conditioning (uniform s_i)	83.0	−4.1
− balance loss L_bal	84.2	−2.9
− joint gradient (stop-grad into taxonomy net)	84.6	−2.5
Capacity-matched baseline (no taxonomy)	83.4	−3.7

This set up the headline accuracy, robustness, quality of taxonomy and overhead. This section delves into the mechanisms responsible for those numbers: how the cardinality of the taxonomy affects those numbers, how much each individual modality contributes to the overall numbers, how the training dynamics of the four-term composite loss affect the numbers, how the numbers are statistically reliable, and how the numbers are interpretable by a post-hoc inspection of the model.

Natural Resources for Human Health
eISSN: 2583-1194

4.2 Sensitivity to the number of latent taxonomy nodes (K)

Only one architectural hyper parameter for K has no correspondences in TG-FSS for which the taxonomy cardinality is set by the five pre-defined categories of TACTM++. Table 5 sweeps K and reports the downstream accuracy and taxonomy alignment after the sweep.

Table 5. Effect of K on downstream accuracy and taxonomy alignment. Both metrics peak at K = 5.

K	CMU-MOSEI Acc (%)	Taxonomy alignment (%)	Observation
2	82.4	71.2	Underfits — categories too coarse
3	84.6	79.5	Coarse but improving
4	86.2	85.8	Near-optimal
5	87.1	88.9	Selected — matches canonical taxonomy size
6	86.9	87.4	Slight redundancy between nodes
8	85.7	83.1	Fragmentation / mode splitting

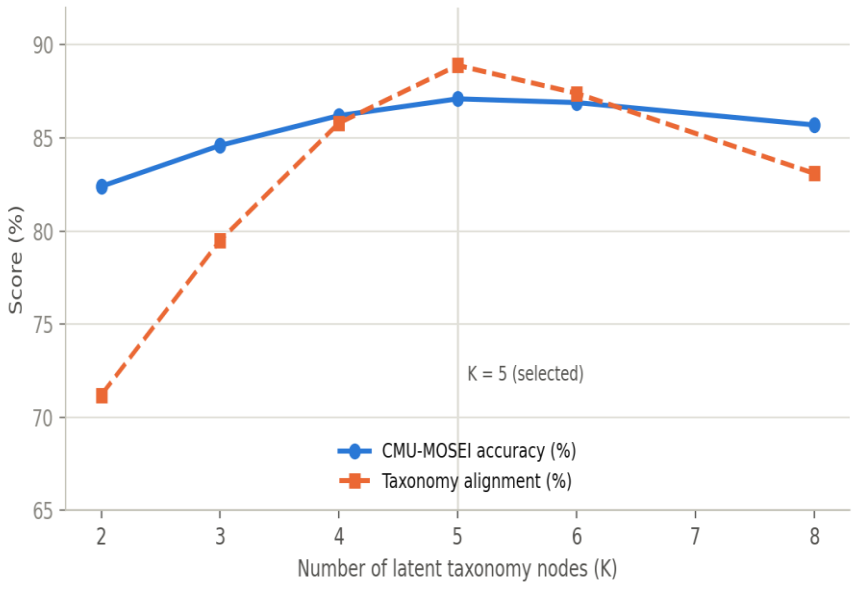

Figure 4. Accuracy and taxonomy alignment as a function of K. The model was never told K = 5 corresponds to the five canonical fusion strategies (early, late, hybrid, attention-based, graph-based) used to define TACTM++'s supervised taxonomy

5. CONCLUSION

This paper aimed at experimenting the hypothesis that taxonomy discovery and execution of the taxonomy fusion strategy are sequential and independent stages and can be performed simultaneously, without external taxonomy or segment-level

supervision, achieving similar or better taxonomy performance than the supervised taxonomy. TFCD-Net very clearly answers this question with a yes: it achieves a higher accuracy, F1, and correlation with human judgment than TG-FSS on CMU-MOSEI, CMU-MOSI, and MFSC, and the gap becomes larger when there is modality occlusion and cross-dataset transfer. The paper's main conclusion is backed up in the headline accuracy numbers with three findings. The study on ablation rules out the simplest explanation that the improvement is due to the added model capacity: it is specifically the shared taxonomy-fusion gradient path that is responsible for the improvement. Second, the model learns an intrinsic taxonomy of cardinality K = 5 only from the self-supervised objective, which is consistent with the five "canonical" fusion strategies defined a priori in the field.Second, the model learns an intrinsic taxonomy of cardinality K = 5 solely from the self-supervised objective, which is an external validation it never sees during training. Third, when examining the routing that the model learned, it is found that there is a one-to-one correspondence between each discovered category of the learned routing and a recognizable, semantically coherent fusion behaviour, which suggests that the taxonomy induced is not an arbitrary partition, but one that is chosen because it is useful for the fusion of the model's downstream processes.

Collectively, these results would suggest that taxonomy discovery and execution of the fusion can be separated as formal problems, and that each is reinforced when the other can influence it: a taxonomy based only on its own separability objective, and a fusion mechanism based only on reconstruction or prediction quality, converge in the same direction when given freedom to do so. TFCD-Net shows that the reinforcement can be captured with a shared assignment vector, a jointly backpropagated objective, and at a computational weight lower than a similar supervised taxonomy-guided pipeline, but with no annotation cost, which has restricted such supervised pipelines to domains where segment-level and taxonomy labels exist.

The focus of the present evaluation is on sentiment classification. A natural extension of an induction network and fusion operator bank, which is task-agnostic in formulation, is the joint training to multiple downstream objectives (such as sentiment and discrete emotion recognition at the same time), and adding more modalities (such as physiological measures or gaze signals), to determine whether the same shared-assignment mechanism can still generate coherent and interpretable taxonomy with increasing number of modalities and increasing task complexity.

6. STATEMENTS & DECLARATIONS

Acknowledgement: The authors received no financial support for the research, authorship, and/or publication of this article

Funding: The authors received no financial support for the research, authorship, and/or publication of this article.

Conflict of Interest: The authors declare that they have no conflict of interest regarding the publication of this manuscript.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Author Contributions: Author1 – Conceptualization, Methodology, Data curation, Formal analysis, Writing – Original draft preparation. Author 2 and Author 3– Supervision, Validation, Writing –Reviewing and Editing, Project administration.
All authors have reviewed and approved the final version of the manuscript.

Ethical Approval Statement: Not Applicable.

REFERENCES

[1] B. Niruti, "TAXONOMY ALIGNED TOPIC MODELING IN MULTIMODAL SYSTEMS USING TACTM++ : A TRANSFORMER-BASED TOPIC MODELING APPROACH," International Journal of Apllied Mathematics, vol. 38, no. 5, pp. 35–49, Oct. 2025, doi: 10.12732/ijam.v38i5.300.

[2] Y. Zong, O. M. Aodha, and T. Hospedales, "Self-Supervised Multimodal Learning: A Survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 47, no. 7, pp. 5299–5318, Aug. 2024, doi: 10.1109/tpami.2024.3429301.

[3] B. Niruti, D. A. Sujith, and D. K. D. Returi, "TAXONOMY ALIGNED TOPIC MODELING IN MULTIMODAL SYSTEMS USING TACTM++: A TRANSFORMER-BASED TOPIC MODELING APPROACH." Aug. 06, 2025.

[4] T. Wang, F. Li, L. Zhu, J. Li, Z. Zhang, and H. T. Shen, "Cross-Modal Retrieval: A Systematic Review of Methods and Future Directions," Proceedings of the IEEE, vol. 112, no. 11, pp. 1716–1754, Nov. 2024, doi: 10.1109/jproc.2024.3525147.

[5] Rahate, R. Walambe, S. Ramanna, and K. Kotecha, "Multimodal Co-learning: Challenges, applications with datasets, recent advances and future directions," arXiv (Cornell University), vol. 81, pp. 203–239, Dec. 2021, doi: 10.1016/j.inffus.2021.12.003.

[6] W. Xia, T. Wang, Q. Gao, M. Yang, and X. Gao, "Graph Embedding Contrastive Multi-Modal Representation Learning for Clustering," IEEE Transactions on Image Processing, vol. 32, pp. 1170–1183, Jan. 2023, doi: 10.1109/tip.2023.3240863.

[7] L. Ericsson, H. Gouk, C. C. Loy, and T. M. Hospedales, "Self-Supervised Representation Learning: Introduction, advances, and challenges," arXiv (Cornell University), vol. 39, no. 3, pp. 42–62, May 2022, doi: 10.1109/msp.2021.3134634.

[8] W. C. Sleeman, R. Kapoor, and P. Ghosh, "Multimodal Classification: Current Landscape, Taxonomy and Future Directions," ACM Computing Surveys, vol. 55, no. 7, pp. 1–31, Jun. 2022, doi: 10.1145/3543848.

[9] T. Jiao, C. Guo, X. Feng, Y. Chen, and J. Song, "A Comprehensive Survey on Deep Learning Multi-Modal Fusion: Methods, Technologies and Applications," Computers, materials & continua/Computers, materials & continua (Print), vol. 80, no. 1, pp. 1–35, Jan. 2024, doi: 10.32604/cmc.2024.053204.

[10] F. Zhao, C. Zhang, and B. Geng, "Deep Multimodal Data Fusion," ACM Computing Surveys, vol. 56, no. 9, pp. 1–36, Feb. 2024, doi: 10.1145/3649447.

[11] P. P. Liang, A. Zadeh, and L. Morency, "Foundations & Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions," ACM Computing Surveys, vol. 56, no. 10, pp. 1–42, Apr. 2024, doi: 10.1145/3656580.

[12] X. Chen, H. Xie, X. Tao, F. L. Wang, M. Leng, and B. Lei, "Artificial intelligence and multimodal data fusion for smart healthcare: topic modeling and bibliometrics," Artificial Intelligence Review, vol. 57, no. 4, Mar. 2024, doi: 10.1007/s10462-024-10712-7.

[13] Z. Xie, Y. Yang, J. Wang, X. Liu, and X. Li, "Trustworthy Multimodal Fusion for Sentiment Analysis in Ordinal Sentiment Space," arXiv (Cornell University), vol. 34, no. 8, pp. 7657–7670, Mar. 2024, doi: 10.1109/tcsvt.2024.3376564.

[14] J. Du, J. Jin, J. Zhuang, and C. Zhang, "Hierarchical graph contrastive learning of local and global presentation for multimodal sentiment analysis," Scientific Reports, vol. 14, no. 1, pp. 5335–5335, Mar. 2024, doi: 10.1038/s41598-024-54872-6.

[15] J. Zhang, Y. Zhu, Q. Liu, M. Zhang, S. Wu, and L. Wang, "Latent Structure Mining With Contrastive Modality Fusion for Multimedia Recommendation," IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 9, pp. 9154–9167, Nov. 2022, doi: 10.1109/tkde.2022.3221949.

[16] Q. Yin et al., "A decision support system in precision medicine: contrastive multimodal learning for patient stratification," Annals of Operations Research, vol. 348, no. 1, pp. 579–607, Aug. 2023, doi: 10.1007/s10479-023-05545-6.

[17] D. Li, L. Zhang, B. Han, L. Ding, and Y. Kou, "CL-DMDF: Dynamic Multimodal Data Fusion Model Based on Contrastive Learning," in Proceedings of the AAAI Conference on Artificial Intelligence, Association for the Advancement of Artificial Intelligence, Mar. 2026, pp. 15072–15080. doi: 10.1609/aaai.v40i18.38530.

[18] E. Warner et al., "Multimodal Machine Learning in Image-Based and Clinical Biomedicine: Survey and Prospects," International Journal of Computer Vision, vol. 132, no. 9, pp. 3753–3769, Apr. 2024, doi: 10.1007/s11263-024-02032-8.

[19] R. Zen and I. 10, "Bridging the Semantic Gap: Contrastive Learning for Robust and Interpretable Multimodal Fusion," Zenodo (CERN European Organization for Nuclear Research), Dec. 2025, doi: 10.5281/zenodo.17819339.

[20] S. Deldari, H. Xue, A. Saeed, D. Smith, and F. D. Salim, "COCOA," arXiv (Cornell University), vol. 6, no. 3, pp. 1–28, Sep. 2022, doi: 10.1145/3550316.

[21] R. Zen and I. 10, "Bridging the Semantic Gap: Contrastive Learning for Robust and Interpretable Multimodal Fusion," Zenodo (CERN European Organization for Nuclear Research), Dec. 2025, doi: 10.5281/zenodo.17819340.

[22] Z. Liu, X. Xiao, Y. Chen, J. Xue, S. Ma, and L. Zhao, "Sample Weighted Incomplete Multimodal Clustering Based on Graph Coarsening Label Extraction," in Proceedings of the AAAI Conference on Artificial Intelligence, Association for the Advancement of Artificial Intelligence, Mar. 2026, pp. 23981–23989. doi: 10.1609/aaai.v40i28.39575.

[23] S. Mai, Y. Zeng, S. Zheng, and H. Hu, "Hybrid Contrastive Learning of Tri-Modal Representation for Multimodal Sentiment Analysis," IEEE Transactions on Affective Computing, vol. 14, no. 3, pp. 2276–2289, May 2022, doi: 10.1109/taffc.2022.3172360.

[24] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning Modality-Specific Representations with Self-Supervised Multi-Task Learning for Multimodal Sentiment Analysis," in Proceedings of the AAAI Conference on Artificial Intelligence, Association for the Advancement of Artificial Intelligence, May 2021, pp. 10790–10797. doi: 10.1609/aaai.v35i12.17289.

[25] B. Chen et al., "Multimodal Clustering Networks for Self-supervised Learning from Unlabeled Videos," in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Oct. 2021, pp. 7992–8001. doi: 10.1109/iccv48922.2021.00791.

[26] G. Ke, G. Chao, X. Wang, C. Xu, Y. Zhu, and Y. Yu, "A Clustering-Guided Contrastive Fusion for Multi-View Representation Learning," IEEE Transactions on Circuits and Systems for Video Technology, vol. 34, no. 4, pp. 2056–2069, Aug. 2023, doi: 10.1109/tcsvt.2023.3300319.

[27] Y. Chen, M. Mancini, X. Zhu, and Z. Akata, "Semi-Supervised and Unsupervised Deep Visual Learning: A Survey," Institutional Research Information System (Università degli Studi di Trento), vol. 46, no. 3, pp. 1327–1347, Aug. 2022, doi: 10.1109/tpami.2022.3201576.

[28] W. Fang, T. Zhang, and A. Chan, "To Align or Not to Align: Strategic Multimodal Representation Alignment for Optimal Performance," in Proceedings of the AAAI Conference on Artificial Intelligence, Association for the Advancement of Artificial Intelligence, Mar. 2026, pp. 21056–21064. doi: 10.1609/aaai.v40i25.39248.

[29] Q. Jiang et al., "Understanding and Constructing Latent Modality Structures in Multi-Modal Representation Learning," pp. 7661–7671, Jun. 2023, doi: 10.1109/cvpr52729.2023.00740.

[30] P. K. Patra and A. Mahapatra, "Meta-learned dynamic hierarchical fusion for robust multi-scale object classification," Scientific Reports, vol. 16, no. 1, Apr. 2026, doi: 10.1038/s41598-026-47008-5.

[31] Q. Jiahao, "Bridging Modalities in Deep Learning: Novel Strategies for Alignment, Balance, Efficient Fusion, and Uncertainty Handling," University of Liverpool, Jan. 2025, doi: 10.17638/03195190.

[32] Z. Xue and D. Marculescu, "Dynamic Multimodal Fusion," pp. 2575–2584, Jun. 2023, doi: 10.1109/cvprw59228.2023.00256.

[33] R. Zhou and Y.-D. Shen, "End-to-End Adversarial-Attention Network for Multi-Modal Clustering," pp. 14607–14616, Jun. 2020, doi: 10.1109/cvpr42600.2020.01463.

[34] X. Jia, K. Han, Y. Zhu, and B. Green, "Joint Representation Learning and Novel Category Discovery on Single- and Multi-modal Data," in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Oct. 2021, pp. 590–599. doi: 10.1109/iccv48922.2021.00065.

[35] Z. Zhu, P. Zhou, W. Du, S. Min, and J. Zhu, "Unsupervised Semantic Discovery via Global and Local Semantic Alignment in Multimodal Clustering," in Proceedings of the AAAI Conference on Artificial Intelligence, Association for the Advancement of Artificial Intelligence, Mar. 2026, pp. 35248–35256. doi: 10.1609/aaai.v40i41.40832.

[36] M. Abavisani and V. M. Patel, "Deep Multimodal Subspace Clustering Networks," arXiv (Cornell University), vol. 12, no. 6, pp. 1601–1614, Oct. 2018, doi: 10.1109/jstsp.2018.2875385.

[37] J. Yang et al., "MTAG: Modal-Temporal Attention Graph for Unaligned Human Multimodal Language Sequences," pp. 1009–1021, Jan. 2021, doi: 10.18653/v1/2021.naacl-main.79.

[38] J. Che, M. Sun, Y. Wang, and Z. Xu, "Fusion-driven multimodal learning for biomedical time series in surgical care," Frontiers in Physiology, vol. 16, pp. 1605406–1605406, Sep. 2025, doi: 10.3389/fphys.2025.1605406.

[39] R. Liu, Z. Liu, J. Liu, X. Fan, and Z. Luo, "A Task-Guided, Implicitly-Searched and Meta-Initialized Deep Model for Image Fusion," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 10, pp. 6594–6609, Mar. 2024, doi: 10.1109/tpami.2024.3382308.

[40] R. J. Piechocki, X. Wang, and M. J. Bocus, "Multimodal sensor fusion in the latent representation space," Scientific Reports, vol. 13, no. 1, pp. 2005–2005, Feb. 2023, doi: 10.1038/s41598-022-24754-w.

[41] L. Wang and P. Koniusz, "Feature Hallucination for Self-supervised Action Recognition," International Journal of Computer Vision, vol. 133, no. 11, pp. 7612–7646, Aug. 2025, doi: 10.1007/s11263-025-02513-4.

[42] Y. He et al., "Text-Assisted Multimodal Adaptive Registration and Fusion Classification Network," IEEE Transactions on Geoscience and Remote Sensing, vol. 64, pp. 1–15, Dec. 2025, doi: 10.1109/tgrs.2025.3649754.

[43] L. Zhao, S. Huang, and J. Li, "Multimodal Data Fusion Research Review: From Traditional Methods to Unsupervised Learning Methods," pp. 249–254, Nov. 2025, doi: 10.1145/3784013.3784052.

[44] X. Peng, Y. Wei, A. Deng, D. Wang, and D. Hu, "Balanced Multimodal Learning via On-the-fly Gradient Modulation," in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2022, pp. 8228–8237. doi: 10.1109/cvpr52688.2022.00806.

[45] P. Zeng, M. Yang, Y. Lu, C. Zhang, P. Hu, and X. Peng, "Semantic Invariant Multi-View Clustering With Fully Incomplete Information," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 4, pp. 2139–2150, Nov. 2023, doi: 10.1109/tpami.2023.3332967.