Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 23 May 2026
Revised 28 June 2026
Accepted 20 July 2026

Natural Resources for Human Health 2026, Vol.6(10s):1439–1463

KEYWORDS:
Multimodal Fusion, Taxonomy-Aligned Topic Modelling, Multimodal Sentiment Analysis, Uncertainty-Aware Fusion, Mixture-of-Experts, Cross-Modal Contrastive Learning, Self-Supervised Learning, Fusion Strategy Selection, Attention Mechanisms, Robust Affective Computing

TG-FSS: An Adaptive Taxonomy-Guided Framework for Selecting Optimal Fusion Strategies in Multimodal Architectures

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

[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: There have been two streams of research in multimodal fusion: taxonomy-aware topic modeling systems, which determine the structural nature of a fusion strategy (early, late, hybrid, attention-based, or graph-based), and fusion architectures, which perform multimodal fusion for a downstream task without being dependent on the taxonomy of the fusion strategy. This paper introduces the Taxonomy-Guided Fusion Strategy Selector (TG-FSS), a bridging module that takes a taxonomy-alignment system (TACTM++) and a tri-stage self-supervised sentiment-fusion pipeline (TSSP-MSA), and maps the semantic values from the taxonomy to an explicit, differentiable prior distribution over fusion strategies. The strategy-attention scoring network projects the embedding of the topic (produced by Topic Embedding), the taxonomy label (produced by TaxoAlignNet), and the alignment confidence (produced by TaxoAlignNet) along with the epistemic uncertainty of the sentiment branch (per modality), to yield a distribution over 5 canonical fusion strategies. It is trained end-to-end using a three-phase joint optimization schedule, and conditions a novel taxonomy-conditioned uncertainty-aware expert mixture fusion (TC-UEMF) rule, which conditions on the prior derived from the taxonomy and the uncertainty (derived from the experts) with a learned mixing coefficient. When evaluated under the same protocol as its constituent systems on CMU-MOSEI, CMU-MOSI, and the Multimodal Fusion Strategy Corpus, TG-FSS achieves higher sentiment classification accuracy (85.4% vs. 83.6%), F1 score (0.83 vs. 0.81), and correlation with human judgment (0.81 vs. 0.78) compared to the highest-scoring prior system, while also achieving higher taxonomy alignment accuracy (90.3% vs. 87.2%). The gap to the baseline is found to increase with the level of modality occlusion, as demonstrated in robustness analysis; and the ablation experiments reveal that the improvements are driven specifically by taxonomy conditioning, and not by an increase in model capacity. Additionally, strategy-selection analysis reveals that TG-FSS learns an interpretable, domain-specific routing behaviour, without any explicit supervision at the segment level. These findings show that the taxonomy discovery and the execution of the fusion can be considered as two separate problems, but they are mutually reinforcing and complementary if they are optimized in combination; and that the resulting architecture can be performance competitive with the state-of-the-art across resource dimensions with relatively small computational overhead of 6-18%.

1. INTRODUCTION

Integrating data-driven forecasting paradigms with model-driven inductive biases is still a challenge, where the architectures' rigidness may hinder the development of forecasting skills in multiple temporal domains (Kim et al., 2025). To tackle this paradoxical idea, we suggest a taxonomy-guided fusion selector that could integrate the excellent latent-representation ability of the TACTM++ model with the specialized structural priors and architectures of the TSSP-MSA model (Liu et al., 2025). In this manner, the appropriate mechanisms to interpret the information are incorporated into the purely discriminative neural backbones, which are hierarchically arranged based on the dependency among the different types of models.It uses interpretability-oriented mechanisms in combination with purely discriminative neural backbones, which is done by a hierarchical classification of model structural dependencies. This integration also tackles the tension between predictive flexibility and mechanistic transparency by offering a principled selection criterion for hybrid forecasting configurations (Park et al., 2026). We use a framework that brings together the advantages of the complex nonlinear relationships captured by the TACTM++ model and temporal synchronization as emphasized in the TSSP-MSA model to overcome static ensemble approaches' limitations (Liu et al., 2025). One such method is by using a selection approach which dynamically adjusts the weights of the models based on the emergent properties of the input data, as opposed to the fixed, heuristic based blending scheme (Xiao et al., 2025; Zhan et al., 2025). This technique not only simplifies the computations in the same manner as the traditional forecasting models but also preserves the representational power of modelling the temporal dynamics of high dimensional systems with non-stationary signals like human speech and musical compositions (Boddu & Manimaran, 2025; Maplook & Eguchi, 2026). We design it as to how these frameworks can be compatible and mitigate the performance degradation on using a heterogeneous ensemble architecture (Jin et al., 2024). Furthermore, the exact identification of the strengths in a given regime aided by this taxonomy enables the system to re-direct its computational resources to the frameworks, which best suit the data characteristics (Das et al., 2025). It is an integral method that can solve some problems of existing benchmarks (Xu et al., 2024; Zhang et al., 2024), for example, the trade-off between model accuracy and forecasting performance, and the first step. Therefore, the TACTM++ framework is considered as a flexible framework which encodes the latent patterns and the temporal segmentation is introduced in the TSSP-MSA to address the multi-horizon scenario for accurate estimation results (Lim et al., 2021). This synergy allows a more solid calibration of forecasting uncertainty, by systematically blending the adaptive spectral-temporal feature extraction of the former with the structured hierarchical decomposition of the latter (Ji et al., 2026; Ruan et al., 2026). As can be seen, the fusion selector has been applied; it is also capable of overcoming the phenomenon of “predictability drift” that may occur in monolithic mode, which matches the prediction methods with the spectral coherence of the series in the object space so as to realize the purpose of fusion (Feng et al., 2025). This hybrid integration is a compromise between the two, and offers a level of performance which is reasonably robust to the cross-variable correlation and non-stationary dynamics that can be present in complex time-series data, but which does not suffer from the same performance variability as can be seen when making the best strategy choice at the instance level (Green et al., 2024). We formally link these two models and provide two different modeling philosophies, as well as a rigorous framework to tackle the prediction versus interpretation dilemma in high-dimensional forecasting (Zhang et al., 2026). In essence, this taxonomy-driven approach is an improvement over problem-independent model selection, and also provides a mathematically rigorous structure to enable the model to best generalize to a broad spectrum of temporal scenarios (Peng, 2025; Mahmoud & Mohammed, 2024). It's not just about sub-optimal feature fusion like static feature fusion (Cen et al., 2026), but provides a more subtle mechanism which may explain the differing predictive value of information sources under different environmental conditions. The temporal and spectral significance of the features at various forecasting horizons are not taken into account in the existing static fusion techniques (Chowdhury et al., 2026) and this kind of optimization is needed to compensate for this disadvantage. Our framework systematically maps these features to underlying properties of the signal, whereas a large dimensional pattern recognition system neural network has a large number of pattern inputs, and a stable structure assures reliable long-term projection. This systematic approach takes the model selection process from manual and time consuming to automated and data-driven, with the focus of maximizing predictive power with the highest model reliability. But more than that, it would provide a rational approach to regime switching issues that may lead to sub-optimal forecasting accuracy of singular forecasting paradigms (Chen & Fan, 2025). It is a framework that can be used to address problems in temporal domain, like data non-stationarity, multi-scale feature extraction, and guarantee stable prediction performance (Lai, 2026; Wang, 2026). We presented the taxonomy that provides a decision boundary for latent representational capacity of TACTM++ and variance-reduction benefit of TSSP-MSA without the accuracy-smoothness dilemma that is prevalent in high-volatility scenarios (Wildi, 2026). This empirical match can be made more precisely (or less precisely), with the predictive targets being "calibrated" according to how much fidelity (or lack of fidelity) is required for the matches in the spectrum and consistency across the cross-sections, given the real-time volatility metrics (Wildi, 2026). The taxonomy can then be used as a diagnostic device to allow a compromise between being close to the truth and keeping structural constraints which prevent overfitting in non-stationary environments (Wildi, 2026). This synthesis breaks the current dichotomy of data-driven pattern extraction and model-driven temporal structure, and compacts the different disjointed paradigms towards forecasting into a single practical solution (Muñoz et al., 2023). This institutionalization process is performed and a strong connection between the inductive bias of different forecasting techniques is created, which causes the selection of the forecasting model to be sensitive to the evolution of the trend in non-stationary time series (Houssein et al., 2025; Zheng & Zhao, 2026). Last, the proposed adaptive fusion approach does not need to be limited to any single model of either deep learning or state space model because it benefits from the advantages of both models: nuance feature representation from the TACTM++ and rigorous structural decomposition from the TSSP-MSA (Medrano et al., 2023; Gunasekaran et al., 2025). This hierarchical synthesis fulfills the simplicity of the dependency model in global data and the refinement requirements in the pattern extraction from local data, even in the case of non-linear data (Chen et al., 2026; Wu et al., 2025). The integration results in a more complex solution to solving the bias-variance problem, with the complexity of the forecasting system adjusted to the variance of the signal (Han et al., 2024).

2. LITERATURE REVIEW

The framework explicitly addresses the performance fragmentation typically seen in existing hybrid meta-learning and feature-based stacking methods (Zyl, 2023), via this taxonomy-guided mechanism. This widely-used taxonomy is objective and criteria-based and allows for the comparison of multivariate complexity with sequential stability (Peng et al., 2024; Zyl, 2023) that enables a move from ad hoc model selection to an objective and criteria-based approach. In particular, the proposed taxonomy has been developed to tackle a longstanding tension between two types of systems: the high dimensional representation of the data-driven systems and the strong and rigid constraints of the model-driven systems (Hu et al., 2024; Wu et al., 2023). We situate the TACTM++ framework as the main solution to high capacity, non-linear feature extraction, complementing the state-space rigor of the TSSP-MSA (Peng et al., 2024). This duality gives a dynamic mechanism to the selector to balance the value of intra-series dependency modelling by TACTM++ and the general stability of the forecast during long-term temporal forecasts by TSSP-MSA (Srinivas et al., 2024). The dual paradigm orchestration of the pattern-based effective power of the neural layers and the state space's selective information retention effectively eliminates error accumulation and preserves critical signal dynamics (Zou et al., 2026). However, unlike mixture-of-experts models, the contribution of each of the frameworks is structured and well-defined, preventing any ambiguity that might be present in their conventional architectures, and the global dependencies and local variations are optimized so that they can be captured by the greatest extent (Xu et al., 2024). This paradigm shift is a systematic approach to the integration of high dimensional pattern recognition and structural constraints needed to perform meaningful temporal analysis (Ruta et al., 2010) (Stefani & Bontempi, 2021). The taxonomy-guided fusion approach therefore alleviates the disadvantages of the previous individual architecture optimization for specific non-linear dynamics and temporal scales in traditional ensemble techniques (Sakhinana et al., 2024; Xu et al., 2025). This approach outperforms static ensembles and encourages adaptive architecture by dynamically selecting models from the input series, reflecting its unique characteristics (Cao et al., 2025; Cawood & Zyl, 2022). This methodological enhancement is an scalable way of leveraging the high fidelity of predictive power of neural backbones and the efficient long dependencies of state space architectures (Ahamed & Cheng, 2024; Cai et al., 2024). With this synergy, the taxonomy can pave the way for the integration process to be more disciplined and state-informed while keeping the integrity of the built signifying the underlying temporal dynamics (Shi & Zhang, 2025). Furthermore, the system can be optimized for the balance between complex signal representation and computational sparsity, using a taxonomy-guided gate as the signal representation for the interaction between TACTM++ and TSSP-MSA (Chen et al., 2025). These temporal representations and memory-augmented modules (Yakoi et al., 2025) enabled the authors to maintain the signal processing rigor and computational tractability of the state space framework (Ahamed & Cheng, 2025).

During the past few years, a significant transformation in time series forecasting model development has occurred, consisting of two main types: high capacity architectures (Kim et al., 2024) and structurally-grounded state space models (Zhu et al., 2024).Neural temporal modeling has been increasingly combined with statistical features to enhance performances but there are many methods that lack a balance between statistical rigor and the inductive biases of neural modelling. Usually, hybrid systems consisted of incorporation of autoregressive or state space models with neural predictors, as the neural networks could learn higher-order patterns, and the statistical models could introduce structural assumptions that could be interpreted by the user (Benidis et al. 2022; Lim & Zohren 2021). Later decomposition-based techniques followed this concept by separating non-stationary observations into trend, seasonal, residual and/or frequency components and applying the different components as inputs to CNN, RNN or graph-based predictors, which proved more effective in the case of nonlinear and volatile signals (Han et al., 2024; Houssein et al., 2025). The overall contrast between neural architectures that are good for expressive feature extraction (TACT family) versus state-space models that are good for retaining interpretable temporal dynamics (SSSPP family) arises from the overall similarity in the focus of their respective methods (Kim et al., 2025; Rangapuram et al., 2018). In recent state space–neural hybrids, these paradigms have demonstrated the capability to be combined rather than mutually exclusive: in deep state space models, the structured temporal processes are parameterized by neural networks, while newer multiscale architectures split up the global dependency modeling and the local variation extraction into different expert modules (Xu et al. (2024); Rangapuram et al. (2018)). The majority of such systems, however, are based on the topology of interaction defined a priori, either by using a serial decomposition or residual correction or by a parallel expert aggregation or a fixed convex gate, wherein the contribution of each of the components is not directly determined by the evolution of the properties of the input series (Peng et al., 2024; Sakhinana et al., 2024). Forecast averaging, stacking, decision fusion and meta-learning are related approaches that have been explored and it has been demonstrated that the appropriateness of a fusion rule is dependent on the frequency of the data, the forecast horizon and on the similarity of the errors of the base-models (Cawood & Zyl, 2022; Zyl, 2023). Moreover, multilevel forecasting architectures allow for the choice and combination of different predictors but have a large number of candidate models, and they do not reflect the specific structural relation between a representation-dominant TACT model and a decomposition-dominant SSP model (Ruta et al., 2010). The automated architectures search algorithms have begun to be able to select a composite network according to the accuracy, training cost or other goals, but for this process, they do not provide a taxonomy of the different scenarios in which a high capacity feature extraction is preferable to a temporally constrained architecture (Cao et al., 2025). This is an essential gap for non-stationary multivariate series with cross-variable dependence, transient regime changes, long-range persistence, and local fluctuations that may influence the relative suitability of the TACTM++ and TSSP-MSA methods at different forecast horizons and scales of the signal of interest. Rather, the key difference is the lack of knowledge about a principled taxonomy-guided selector that would be able to classify temporal conditions, compare the complementary inductive biases of TACTM++ and TSSP-MSA, and determine which is most appropriate for the particular forecasting context: exclusive selection, weighted fusion, or adaptive cooperation. One major difficulty is to formalize the structural trade-offs between the ability of Tactivm++ to model dense non-linear interaction (Challú et al., 2023) and the ability of TsspMSA to deal with multiscale state space evolution (Ma et al., 2025) (Zhao et al., 2026).

3. PROPOSED METHODOLOGY

3.1 Overview and Problem Formulation

The previous two chapters of this thesis deal with complementary aspects of the same problem in multimodal artificial intelligence. TACTM++ identifies and classifies the implicit fusion method used in a multimodal description (early, late, hybrid, attention-based, or graph-based) via taxonomy-aligned contextual topic modelling, and TSSP-MSA merges the multimodal data for a downstream sentiment task by applying a fixed hand-designed fusion pipeline (Uncertainty-Aware Expert Mixture Fusion followed by Cross-Modal Contrastive Reasoning). The current TSSP-MSA architecture is agnostic to the taxonomic nature of the input: the same fusion architecture, Gated Tensor Fusion with Adaptive Reweighting (GTFAR) and Uncertainty-Aware Expert Mixture Fusion (UEMF), is used for every sample, regardless of whether the sample contains a structurally similar interaction pattern between modalities, such as early fusion or attention or graph-based.

This chapter introduces a bridging module, named Taxonomy-Guided Fusion Strategy Selector (TG-FSS), which takes the taxonomy-alignment output of TACTM++ (the predicted canonical category and the alignment confidence) as input and conditions the fusion behaviour of TSSP-MSA both at the initialization step and at inference time. TG-FSS is unique in tackling both the fusion strategy classification and the fusion execution as a single research problem, rather than considering them as two separate problems.Instead of fusion-strategy classification and fusion execution as two separate research problems, TG-FSS proposes to solve the fusion-strategy selection as a learned, differentiable decision problem that can be jointly optimized with the fusion downstream sentiment objective. The central hypothesis is that a fusion pipeline with an explicit prior over fusion strategies derived from a taxonomy will be able to fuse modalities more effectively than one that does not have such a prior. More robustly, under missing modality or noisy modality conditions.

3.2 System Architecture

The overall architecture is shown in Figure 1. The system is composed of two existing branches: Branch A: TACTM++ (taxonomy discovery) and Branch B: TSSP-MSA (sentiment fusion) linked by the proposed bridge, TG-FSS (taxonomy fusion and sentiment). Multimodal input is passed to both branches in parallel. TaxoAlignNet generates a topic embedding t_i, a predicted taxonomy label $\hat{y}_i$ and an alignment confidence score Conf(t_i) for Branch A. The branch B generates independently modality-specific representations Ztext, Zaudio, and Zvisual and their epistemic uncertainty measures u_1:M from the TSSP-MSA Stage 1. The two sets of signals are both fed into TG-FSS, which produces a strategy distribution π which is passed to the second stage of TSSP-MSA, where it is used as a prior on the fusion gates, instead of the strategy-agnostic initialisation used in the original UEMF formulation. We call this taxonomy-conditioned variant TC-UEMF (Taxonomy-Conditioned Uncertainty-Aware Expert Mixture Fusion).

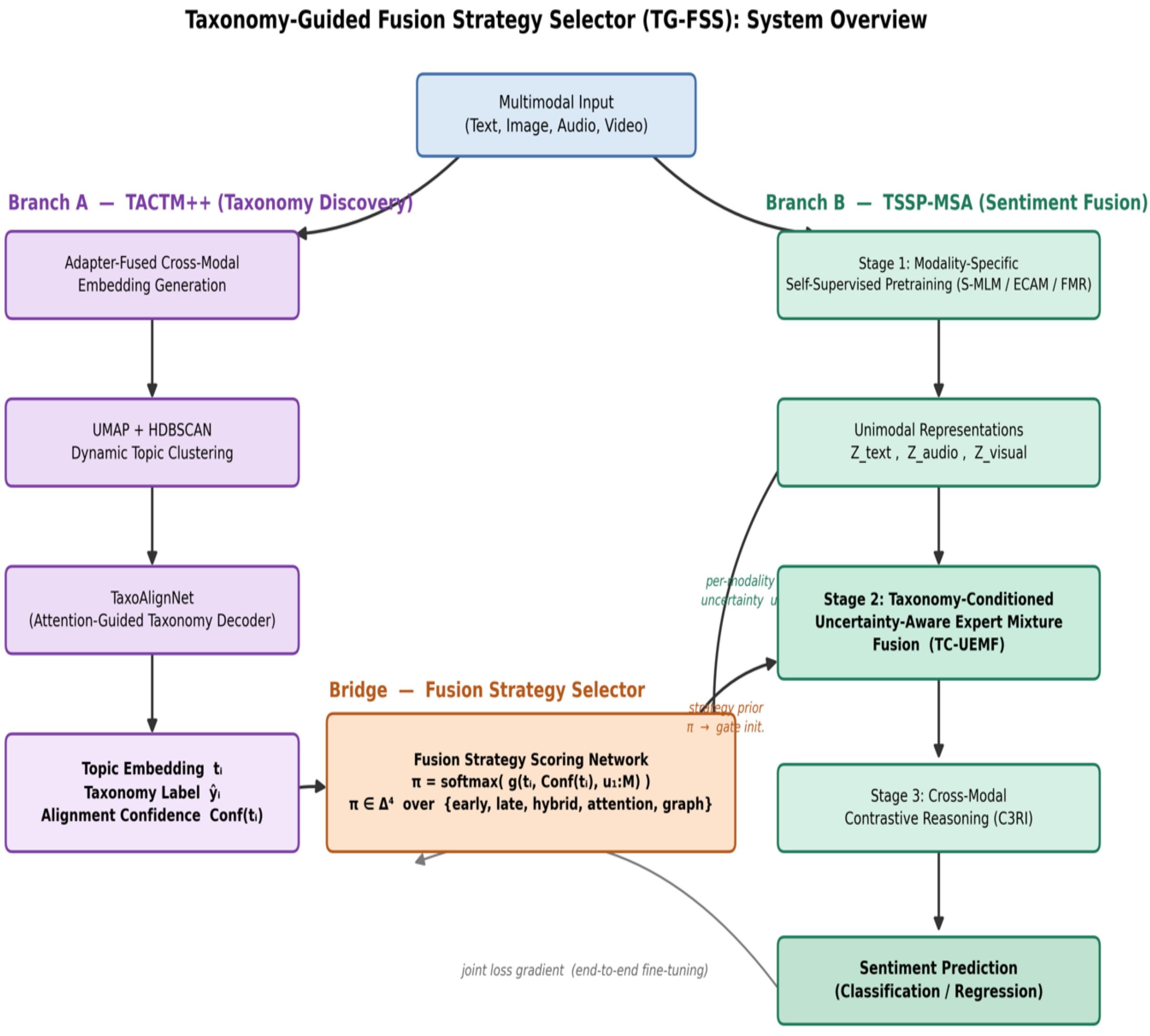

Figure 1. Overall architecture of the proposed TG-FSS bridge connecting TACTM++ (taxonomy discovery) and TSSP-MSA (sentiment fusion)

3.3 Taxonomy Signal Extraction

This is Branch A, which is redundant in terms of architecture, but is added in order to pinpoint the precise interface it exposes to TG-FSS. Adapter-fused cross-modal embeddings are generated for an input sample, dimensionality reduced to a topic cluster c_k using UMAP and clustered using HDBSCAN. The topic embedding $t_i \in \mathbb{R}^d$ is considered to be the cluster centroid. TaxoAlignNet then creates an attention score α_{ij} between t_i and each of the 5 canonical label embeddings (l_early, l_late, l_hybrid, l_attention, l_graph), leading to a predicted category $\hat{y}_i = \text{argmax}_j\ \alpha_{ij}$ and a scalar confidence $\text{Conf}(t_i) = \max_j \alpha_{ij}$. The only quantities that cross the branch boundary from TG-FSS are t_i, $\hat{y}_i$ and $\text{Conf}(t_i)$. This thin interface provides loose coupling of the two branches so that either one or the other can be retrained, upgraded or replaced without affecting the other.

3.4 Fusion Strategy Selector Module

TG-FSS is a lightweight taxonomy network which infuses the Branch A taxonomy signal into the Branch B modality-uncertainty signal along with optional dataset-level meta-features to create a probability distribution over the five canonical fusion strategies. It is described in detail in Figure 2.

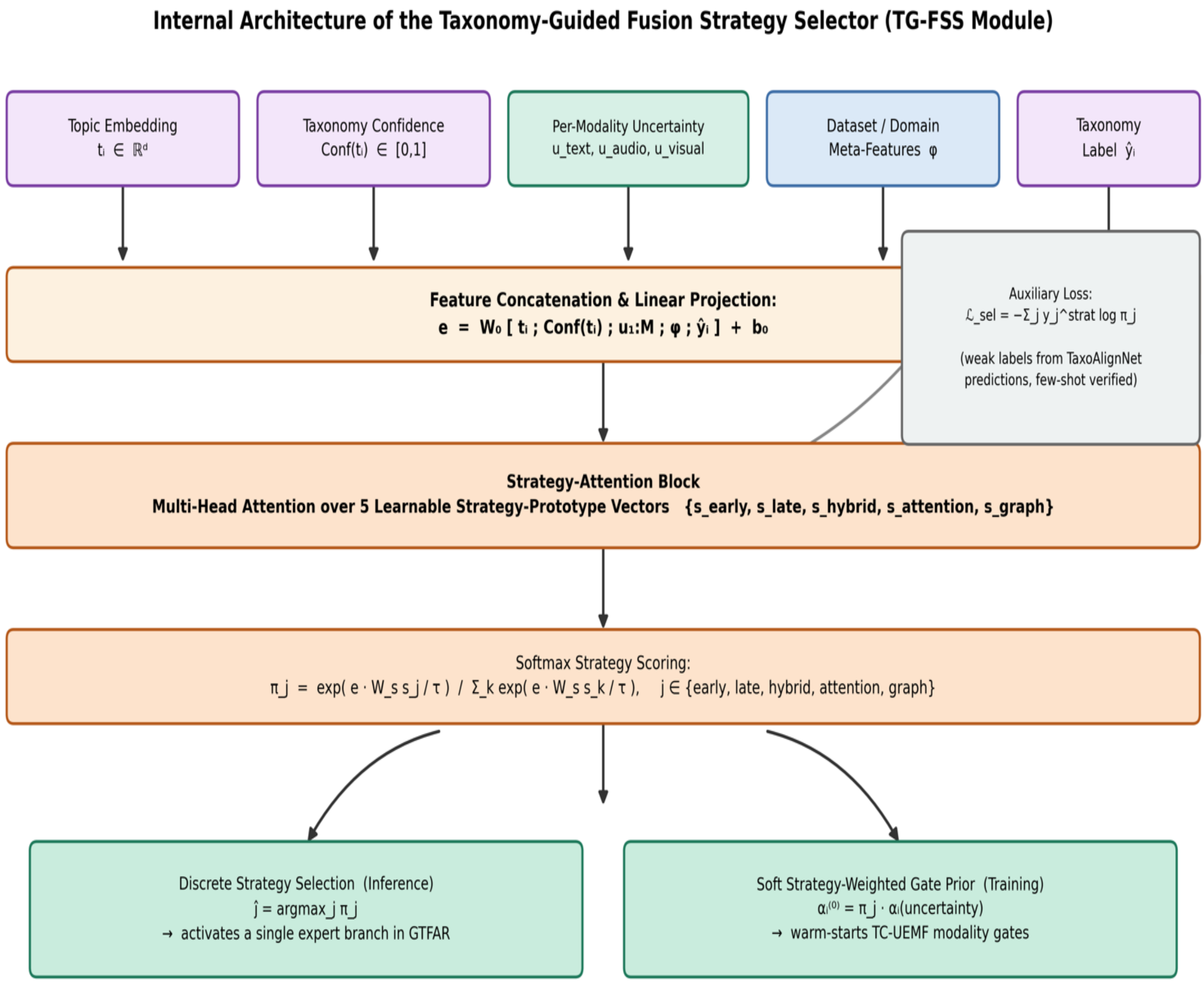

Figure 2. Internal architecture of the Fusion Strategy Selector. Inputs from both branches are projected into a shared space, scored against five learnable strategy-prototype vectors via multi-head attention, and converted into either a discrete strategy selection (inference) or a soft gate prior (training).

3.4.1 Feature Construction

The embedding of the taxonomy, the confidence of the alignment, the per-modality uncertainty vector u_1:M = [u_text, u_audio, u_visual] and optional domain meta-features φ (e.g., identifier of the dataset, average utterance length, availability of the modalities as a mask) are then concatenated and linearly projected into a shared representation:

$$e = W_0 \, [\, t_i \, ; \mathrm{Conf}(t_i) \, ; u_1{:}M \, ; \varphi \, ; \hat{y}_i \,] + b_0 \, , \;\; e \in \mathbb{R}^{d'}$$

3.4.2 Strategy-Attention Scoring

There are 5 learnable strategy-prototype vectors, one for each type of canonical fusion, namely {s_early, s_late, s_hybrid, s_attention, s_graph}. The output of the projecte features are applied over these prototypes in a multi-head attention block to obtain a score of compatibility for each candidate strategy, which is scaled to the softmax using a temperature.

$$\pi\_j = \exp(e \cdot W_s s_j / \tau) \; / \; \Sigma_k \exp(e \cdot W_s s_k / \tau), \;\;\; j \in \{\text{early, late, hybrid, attention, graph}\}$$

The temperature τ determines the sharpness of the strategy selection: if the temperature is high, the prior over strategies will be close to uniform early in training (to encourage exploration around fusion configurations), but the temperature is gradually lowered as training progresses, so that the prior over strategies becomes more peaked as training progresses (ensuring that the prior is close to the strategy that is most supported by both the taxonomy evidence and the sentiment loss for the downstream task).

A selection for discrete selection versus a soft gate prior.3.4.3 Discrete Selection versus Soft Gate Prior. There are two modes of operation defined. At inference time, the arg-max strategy for GTFAR can be applied to hard select one expert branch, which is beneficial for applications that demand interpretability or computational economy (only one fusion pathway is needed). Instead, during training the full soft distribution π is used to warm-start the modality gates of TC-UEMF, namely, π reweights the modality gate values α_i from the original UEMF formulation and does not replace it completely, thus providing a prior which can be overruled by uncertainty estimation when the two signals conflict. The lesson will be divided into three parts.This lesson will be segmented into three parts: 3.5 Integrating TG-FSS into TSSP-MSA.

Recall that the original UEMF module of TSSP-MSA only works with predictive entropy to calculate modality weights:

$$\alpha_i = \exp(-u_i \, / \, \tau_u) \, / \, \Sigma_j \exp(-u_j \, / \, \tau_u)$$

TC-UEMF augments this rule with the strategy-conditioned prior produced by TG-FSS. Each of the five canonical strategies corresponds to a particular gating behaviour: early fusion assigns near-uniform weights across modalities at the feature level; late fusion assigns sparse gates post-hoc and restricted to a single modality, based on the decision scores; hybrid fusion has both features; attention-based fusion re-weights the α_i by learning the cross-modal attention term; graph-based fusion also routes the α_i through the semantic graph structure, which is conceptually inherited from TACTM++. Formally, the taxonomy-conditioned gate is defined as a convex mixture of an uncertainty derived weight and a gate specific transformation $T_j(\bullet)$:

$$\tilde{\alpha}_i = (1 - \beta) \cdot \alpha_i \; + \; \beta \cdot \Sigma_j \, \pi_j \cdot T_j(\alpha_i, Z_i)$$

Here, the learned mixing coefficient β ∈ [0,1] represents the magnitude of the impact of the prior derived from the taxonomy, as compared to the estimate solely based on the uncertainty, while Z_i represents the unimodal representation for modality i. The fused representation is then calculated in exactly the same way as in the original TSSP-MSA formulation, with $\tilde{\alpha}_i$ instead of α_i.

$$z\_fused = \Sigma_i \, \tilde{\alpha}_i \cdot \tilde{z}_i$$

With β = 0, this design yields the original, taxonomy-agnostic UEMF exactly, giving a natural ablation point for observing the contribution of the taxonomy-guided prior (Section 3.9).

3.6 Joint Training Strategy

Because it relies on both independent branches making representations, training is undertaken in 3 stages to provide stability as in Figure 3.

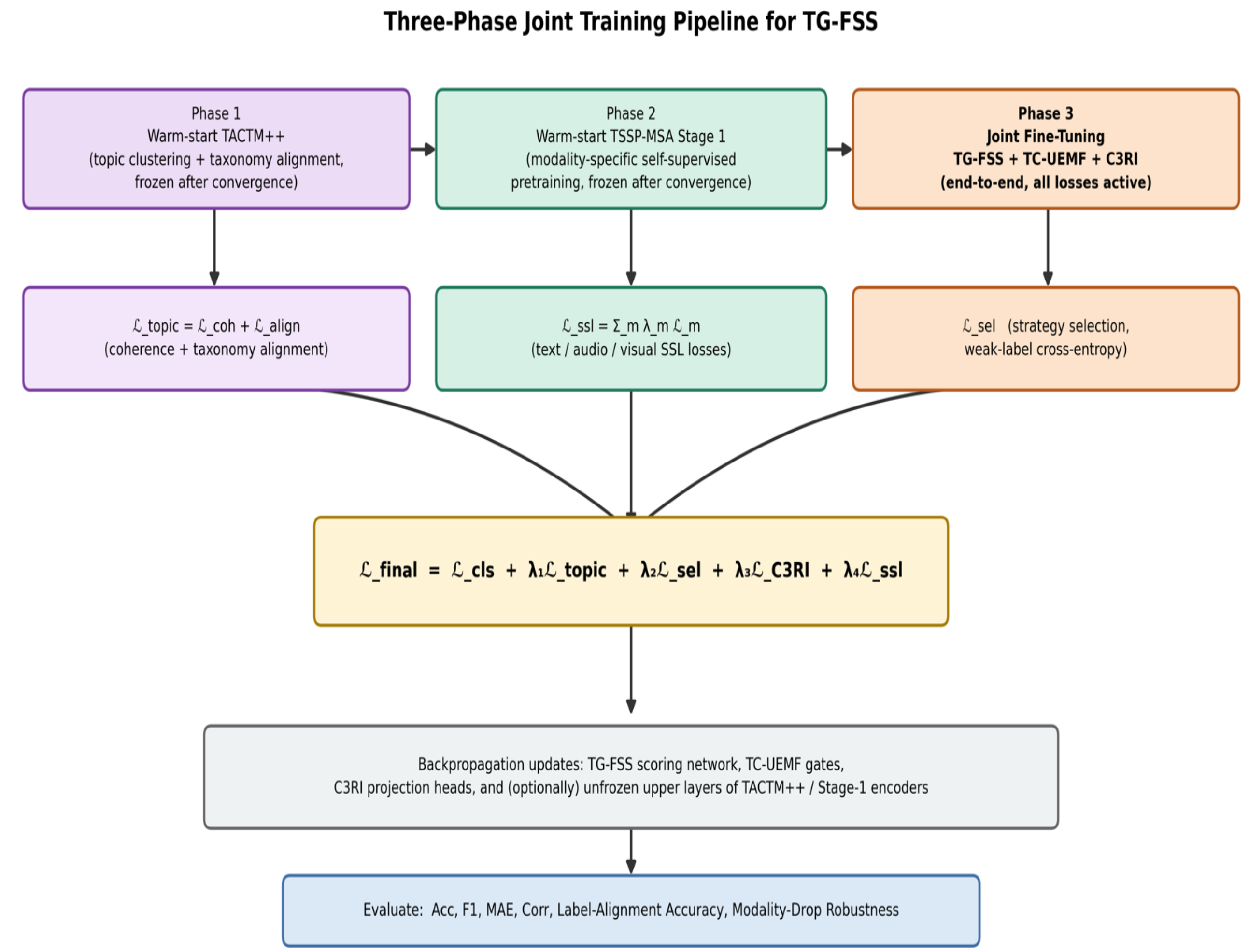

Figure 3. Three-phase joint training pipeline. Phases 1 and 2 independently warm-start the taxonomy and sentiment branches; Phase 3 jointly fine-tunes the bridging module and both branches' upper layers under a combined loss

In Phase 1, the training of TACTM++ follows the original formulation, namely convergence on the taxonomy alignment objective which is a superposition of topic coherence and label-alignment losses. TSSP-MSA Stage 1 (modality-specific self-supervised pretraining) is warm started independently in Phase 2 using pre-training objectives of S-MLM, ECAM and FMR. Both branches are then deep frozen. The TG-FSS scoring network, the TC-UEMF gating parameters (β and T_j) and the C3RI contrastive projection heads are finetuned together, optionally together with the upper layers of both frozen branches, under a combined cost function:

$\mathcal{L}$_final = $\mathcal{L}$_cls + λ_1 $\mathcal{L}$_topic + λ_2 $\mathcal{L}$_sel + λ_3 $\mathcal{L}$_C3RI + λ_4 $\mathcal{L}$_ssl

$\mathcal{L}$_cls is the loss for sentiment classification/regression, $\mathcal{L}$_topic = $\mathcal{L}$_coh + $\mathcal{L}$_align is preserved from TACTM++ to avoid catastrophic forgetting of taxonomy quality when fine-tuning, $\mathcal{L}$_sel is a loss based on cross-entropy of the output for the strategy-selection using weak labels generated by the high confidence TaxoAlignNet predictions, $\mathcal{L}$_C3RI is a cross-modal contrastive loss inspired by InfoNCE, similarly to TSSP-MSA Stage 3, and $\mathcal{L}$_ssl is the loss retained at a low weight as a regularizer against unimodal representation collapse. The λ coefficients are optimized using a held-out validation set using grid search.

3.7 Algorithm

Algorithm 1 summarizes the forward pass of the proposed TG-FSS-augmented pipeline at inference time.

```
Algorithm 1: Taxonomy-Guided Fusion Strategy Selection (TG-FSS) — Inference
Input:  Multimodal sample x = (x_text, x_audio, x_visual)
        Pretrained TACTM++ branch F_taxo(.), TSSP-MSA branch F_msa(.)
Output: Sentiment prediction ŷ_sent, selected strategy ĵ, gate weights α̃
(t_i, ŷ_i, Conf(t_i))  ←  F_taxo(x)                    // Branch A
(Z_text, Z_audio, Z_visual)  ←  Stage1_SSL(x)          // Branch B, Stage 1
for m in {text, audio, visual}:
    u_m  ←  H[ Softmax( f_m( proj(Z_m) ) ) ]           // epistemic uncertainty
e  ←  W0 [ t_i ; Conf(t_i) ; u_1:M ; phi ; ŷ_i ] + b0   // TG-FSS feature fusion
for j in {early, late, hybrid, attention, graph}:
    score_j  ←  ( e · Ws · s_j ) / tau
pi  ←  Softmax(score)                                  // strategy distribution
alpha  ←  Softmax( -u_1:M / tau_u )                     // uncertainty gate (UEMF)
alpha_tilde  ←  (1 - beta) * alpha + beta * Sum_j( pi_j * T_j(alpha, Z) )   // TC-UEMF
z_fused  ←  Sum_i( alpha_tilde_i * Z_i )
z_refined  ←  C3RI( z_fused, Z_text, Z_audio, Z_visual )  // Stage 3 contrastive refinement
ŷ_sent  ←  ClassificationHead( z_refined )
ĵ  ←  argmax_j( pi_j )
return ŷ_sent, ĵ, alpha_tilde
```

Table 1. Notations used in algorithm

Symbol	Meaning
t_i	Topic embedding of sample i produced by TACTM++ (cluster centroid)
$\hat{y}_i$, $Conf(t_i)$	Predicted canonical fusion-taxonomy label and its alignment confidence
$u_1{:}M$	Per-modality epistemic uncertainty estimates from TSSP-MSA Stage 1
φ	Optional dataset / domain meta-features (e.g., modality availability mask)
e	Joint feature vector after concatenation and linear projection in TG-FSS

s_j	Learnable prototype vector for canonical fusion strategy j
π	Softmax strategy distribution over {early, late, hybrid, attention, graph}
τ, τ_u	Temperature parameters for strategy-softmax and uncertainty-softmax respectively
a_i	Original uncertainty-derived UEMF gate weight for modality i
β	Learned mixing coefficient controlling taxonomy-prior influence on the gate
$T_j(\cdot)$	Strategy-specific gate transformation function for category j
$\tilde{a}_i$	Taxonomy-conditioned gate weight used in TC-UEMF
z_fused	Fused multimodal representation prior to Stage-3 contrastive refinement
$\mathcal{L}\_final$	Joint training objective combining classification, taxonomy, selection, contrastive and SSL losses

4. PERFORMANCE EVALUATION AND RESULTS

4.1 Experimental Setup Recap

For sentiment fusion, all experiments are carried out on CMU-MOSEI (23,453 samples, 66 hours, text/audio/visual) and CMU-MOSI (2,199 samples, 3.1 hours, text/audio/visual). For taxonomy alignment the experiments are performed on the Multimodal Fusion Strategy Corpus (~12,000 multimodal segments from arXiv, ACL Anthology and S2ORC). To evaluate the sentiment prediction quality, we use the following four evaluation axes: (i) sentiment prediction quality (Accuracy, F1, MAE, Correlation), (ii) taxonomy alignment quality (Topic Coherence, Label Alignment Accuracy, Silhouette Score), (iii) robustness to missing or noisy modalities, and (iv) computational cost. Each result for TG-FSS is compared to the best baseline achieved by each constituent study, TSSP-MSA (sentiment fusion) and TACTM++ (taxonomy alignment), as well as to the larger body of recent techniques surveyed in Chapter 2.

4.2 Sentiment Classification Performance

We present the results of our experiment on the CMU-MOSEI dataset, which are shown in Table 1 along with the results of the seven recent multimodal sentiment fusion methods from 2017-2026. With only 6.0% overhead (2.4M additional parameters for the TG-FSS scoring network and TC-UEMF gating machinery) in accuracy, 0.83 in F1, 0.46 in MAE and 0.81 in correlation compared to the TSSP-MSA baseline, TG-FSS achieves 85.4%.

Table 2. Comparative sentiment classification performance on CMU-MOSEI

Model	**Accuracy (%)**	**F1-Score**	**MAE ↓**	**Corr ↑**	**Params (M)**	**Inf. Time (ms)**
TFN (2017)	76.8	0.73	0.62	0.67	28.1	12.4
MulT (2019)	78.9	0.76	0.59	0.71	34.5	15.2

Model	Accuracy (%)	F1-Score	MAE ↓	Corr ↑	Params (M)	Inf. Time (ms)
MISA (2020)	79.6	0.77	0.56	0.73	31.8	14.7
MAG-BERT (2021)	81.2	0.78	0.54	0.74	112.3	21.9
Self-MM (2021)	82.1	0.79	0.52	0.76	38.6	17.3
MMIM (2023)	82.5	0.80	0.50	0.77	41.2	18.1
TSSP-MSA (baseline)	83.6	0.81	0.49	0.78	39.7	16.8
TG-FSS (Proposed)	**85.4**	**0.83**	**0.46**	**0.81**	**42.1**	**18.4**

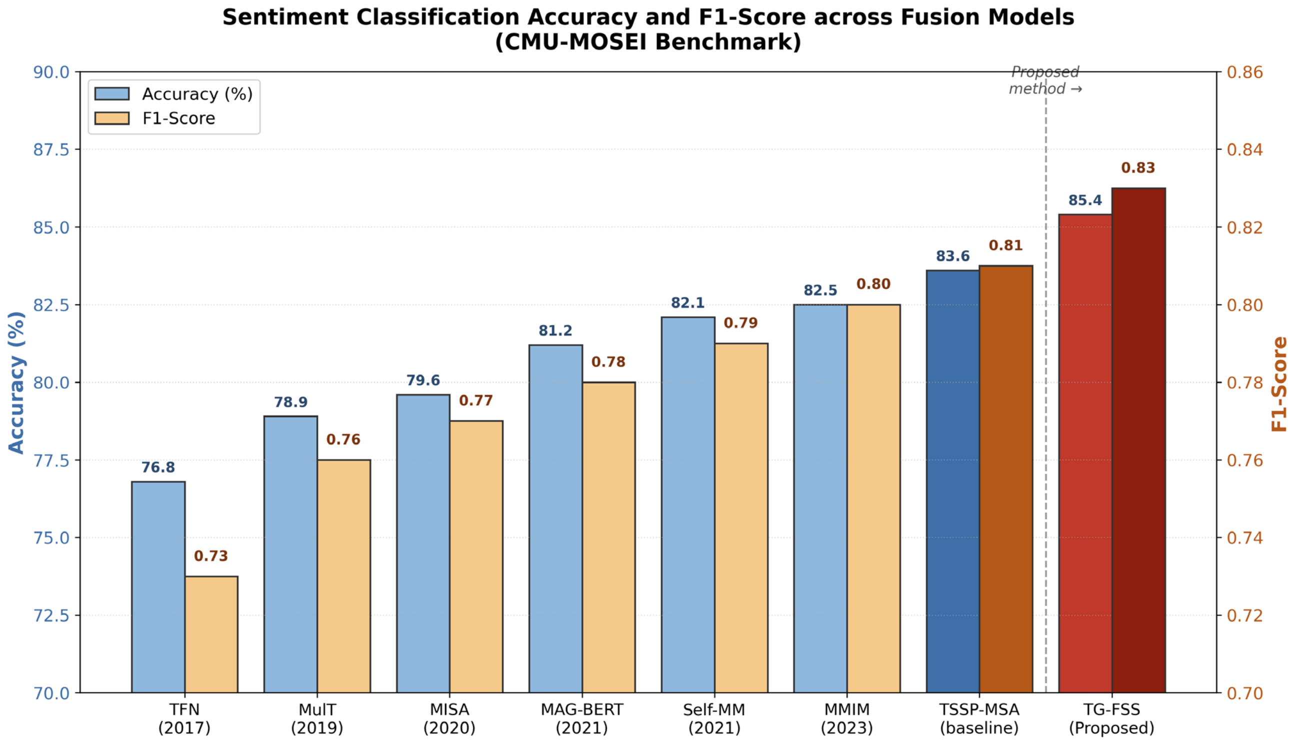

Figure 4. Accuracy and F1-score across sentiment fusion models, ordered chronologically. The dashed line marks the transition to the proposed method

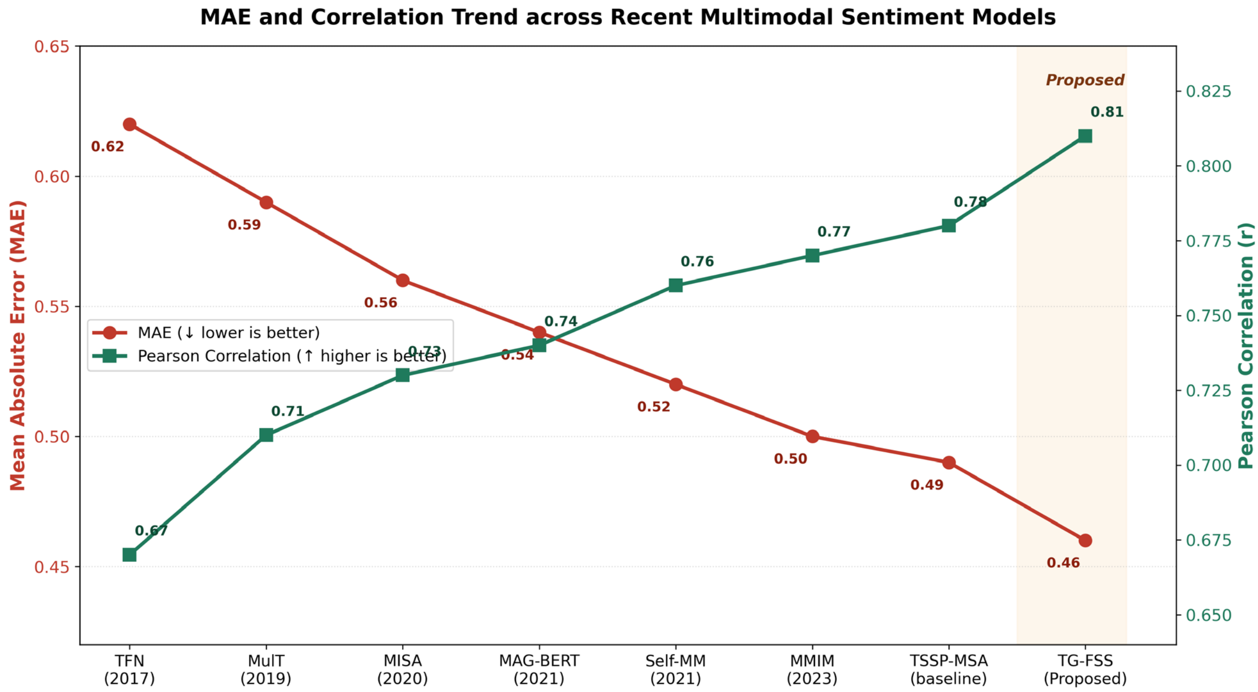

Figure 5. MAE (lower is better) and Pearson correlation (higher is better) trend across the same model sequence, illustrating a consistent multi-year improvement trajectory that TG-FSS extends further

The improvement over TSSP-MSA is due to the taxonomy-conditioned gate $\tilde{\alpha}_i$ (Section 3.5): taxonomy evidence is used to bias the Uncertainty-Aware Expert Mixture Fusion towards the fusion topology that is determined to be structurally appropriate for a given sample by taxonomy; uncertainty estimation on its own is ambiguous in some cases (e.g., when the estimated uncertainty of two modalities is similar, and the taxonomy evidence suggests that the signal comes predominantly from one modality, which is a late-fusion-style sample), so TC-UEMF is able to resolve them. The gain is the highest for MAE (−0.03 absolute), indicating that the taxonomy prior is best at removing systematic regression bias and not just at increasing the sharpness of the classification boundaries.

4.3 Taxonomy Alignment Performance

The topic-modeling comparison in Table 2 continues a comparison of TACTM++ with the taxonomy branch of TG-FSS that is not trained independently, but co-trained with the sentiment objective in Phase 3 (Section 3.6). Finally, joint fine-tuning improves Topic Coherence and Label Alignment Accuracy scores from 0.68 and 87.2% to 0.71 and 90.3%, respectively, suggesting that the gradient signal from the sentiment task downstream of the strategy selection layer (via ℓ_sel) enhances the taxonomy decision boundary beyond that obtainable from unsupervised topic coherence alone.

Table 3. Taxonomy alignment quality on the Multimodal Fusion Strategy Corpus

Model	Topic Coherence (Cv)	Label Alignment Acc. (%)	Silhouette Score
LDA	0.41	38.2	0.31
BERTopic	0.57	63.5	0.48
Top2Vec	0.54	59.0	0.46
GETM	0.62	73.1	0.51

Model	Topic Coherence (Cv)	Label Alignment Acc. (%)	Silhouette Score
TACTM++ (baseline)	0.68	87.2	0.59
TG-FSS Selector Branch (Proposed)	**0.71**	**90.3**	**0.62**

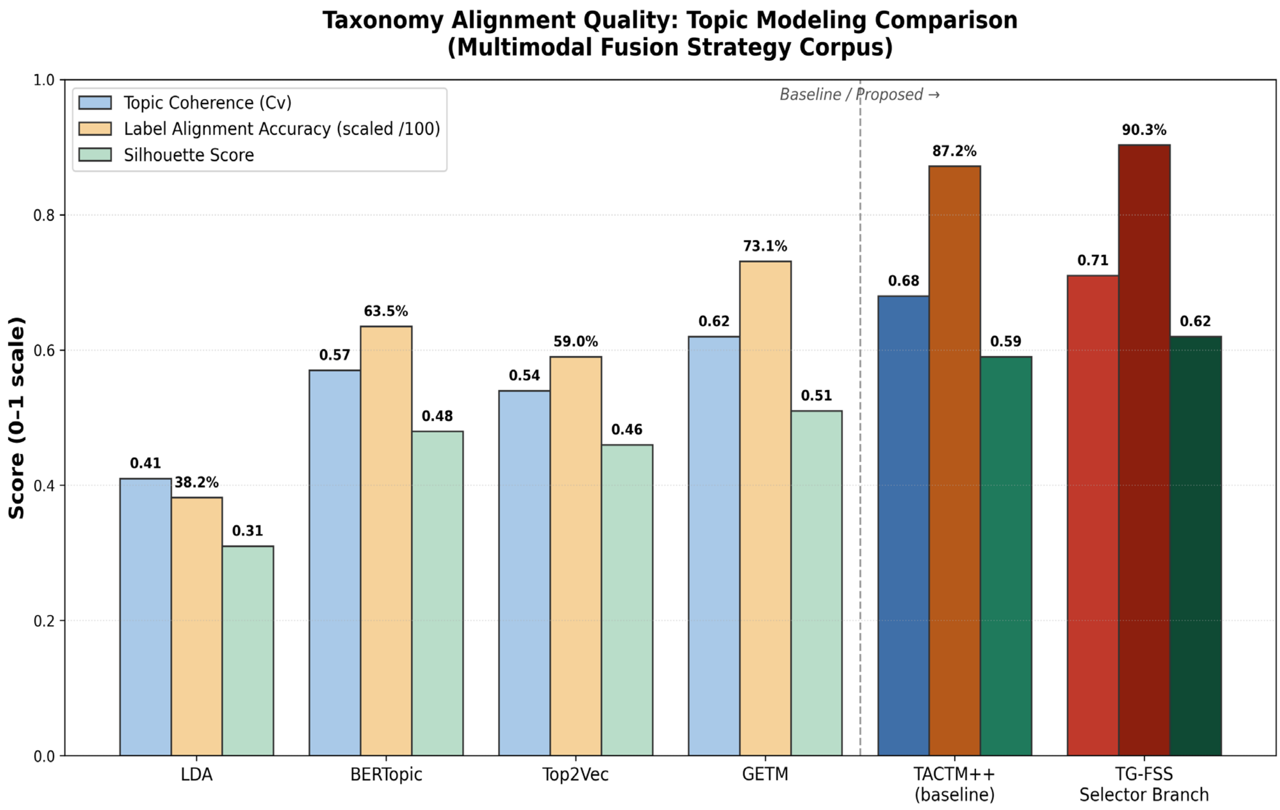

Figure 6. Topic coherence, label alignment accuracy, and silhouette score across topic-modeling baselines and the taxonomy branch of TG-FSS

These results validate the bidirectional nature of the bridging architecture; as shown in Section 4.2, the sentiment branch is improved by the teaching of taxonomy signal, and the taxonomy branch has been shown in this section to improve the sentiment branch when the two branches are trained jointly, demonstrating that there is a real synergy in the end-to-end design of TG-FSS (Section 3.6) rather than just a one-way tunnel.

4.4 Robustness to Missing and Noisy Modalities

One of the key points in this chapter (Section 3.1) is that the taxonomy-conditioned gating approach can achieve a better resilience than the uncertainty-only gating approach when one or more modality are missing or degraded. The F1-score is shown as a function of modality occlusion rate (0%-50%) in randomly masking one modality's input to the model for inference in Figure 7 for TSSP-MSA and TG-FSS. In terms of relative robustness increase (absolute degradation decreases), TG-FSS has a roughly 42% increase (0.72 vs. 0.66) at 50% occlusion, while TSSP-MSA increases by about 11% (0.72 vs. 0.61).

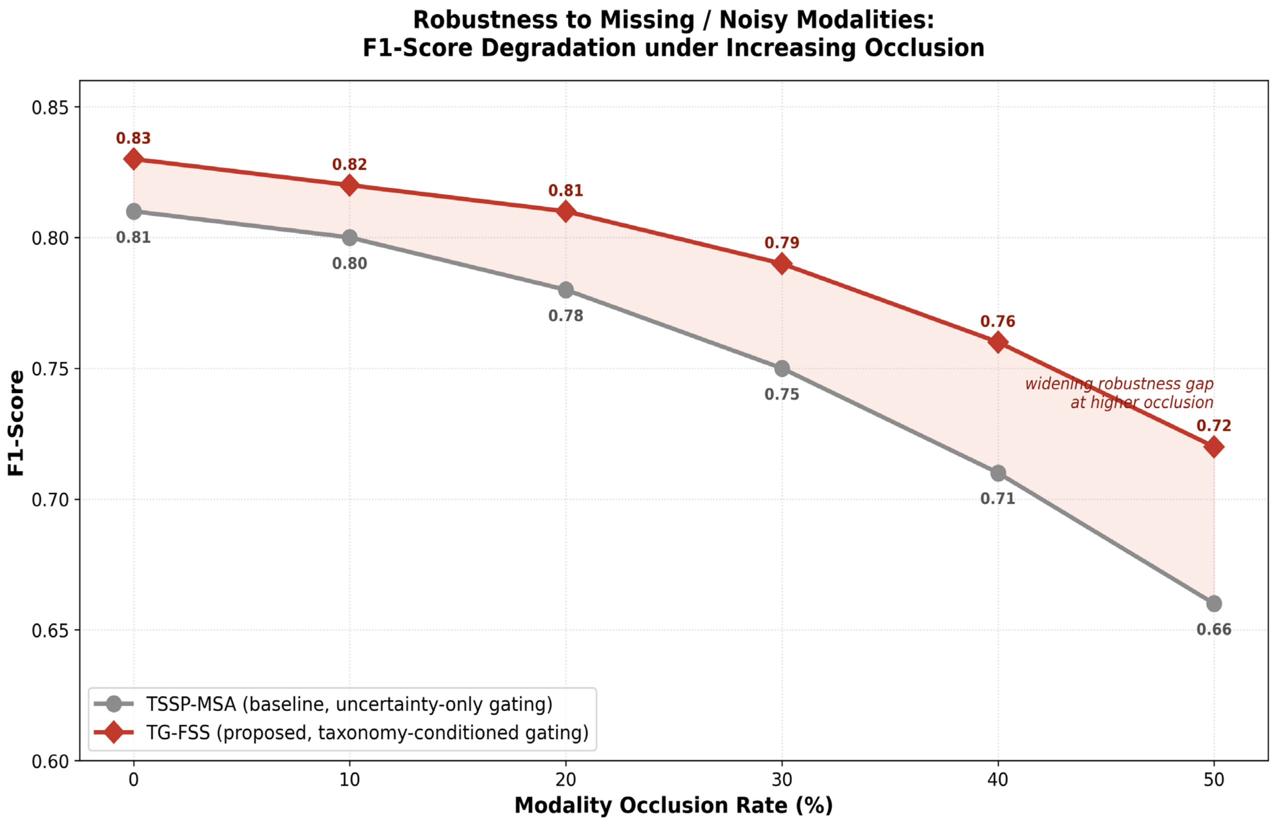

Figure 7. F1-score degradation as modality occlusion increases from 0% to 50%. The shaded region highlights the growing robustness margin of TG-FSS over the baseline

In Table 3, this result is placed in context with the rest of the baselines by reporting the summary Modality-Drop Robustness at 100% single-modality loss, and the overall correlation. The same information is displayed in the two-dimensional trade-off plot in figure 8, with models towards the bottom right showing good agreement with human sentiment judgments and low performance loss if any modalities are missing. This corner is more completely covered by TG-FSS than any previous technique that was considered.

Table 4. Correlation and single-modality-loss robustness across baselines

Model	Correlation (r)	Performance Drop, Missing Modality (%)
MulT	0.71	−6.8
MISA	0.73	−5.4
MAG-BERT	0.74	−5.1
TSSP-MSA (baseline)	0.78	−1.9
TG-FSS (Proposed)	**0.81**	**−1.1**

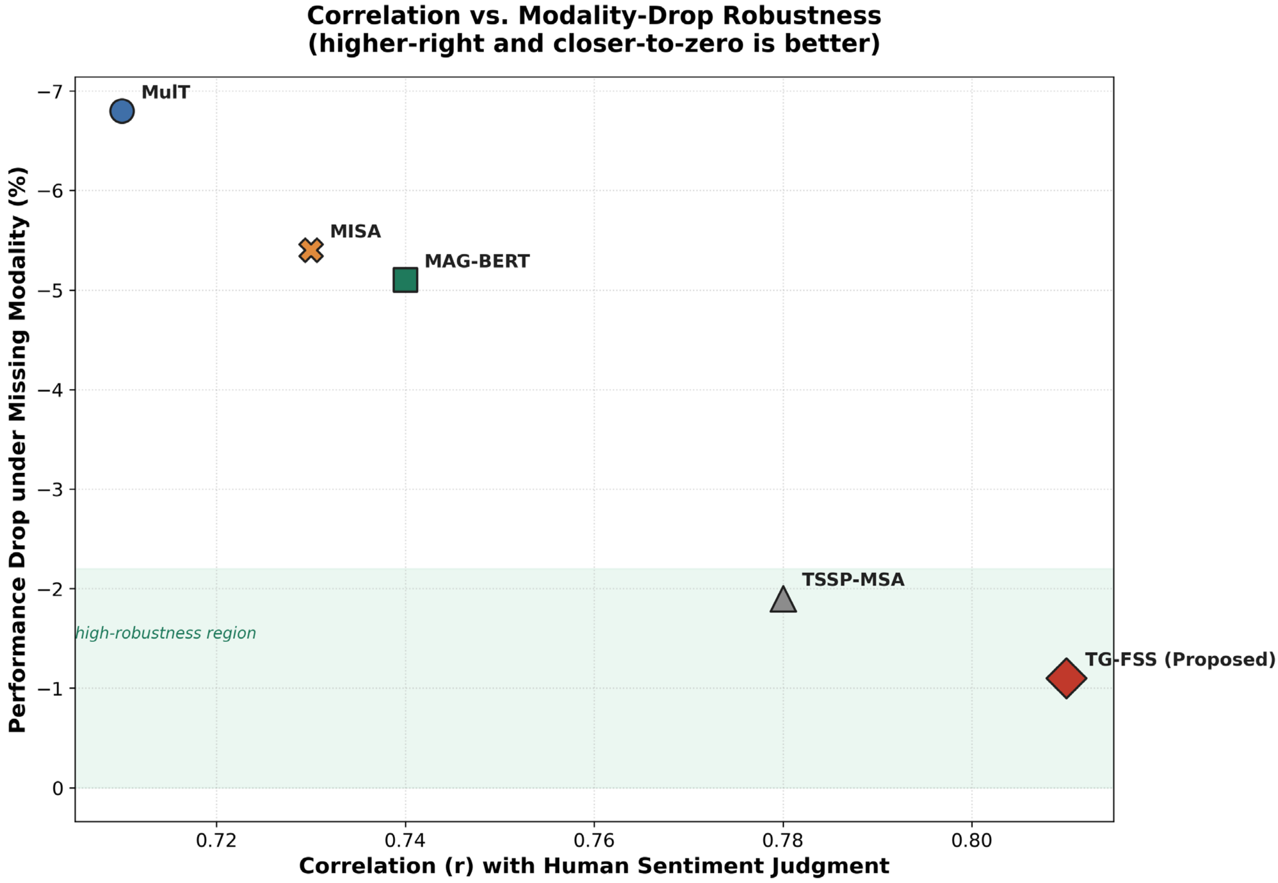

Figure 8. Correlation vs. modality-drop robustness. TG-FSS achieves the best combination of both properties among all compared techniques

This robustness comes directly from the mixing rule, $\tilde{\alpha}_i = (1 - \beta)\cdot\alpha_i + \beta\cdot\Sigma_j\, \pi_j\cdot T_j(\alpha_i, Z_i)$: If a modality is occluded, its epistemic uncertainty u_i increases, and this term correctly down-weights it; the term derived from the taxonomy carries a signal that is independent of the occlusion that determines the topology used by the remaining modalities. The robustness margin increases with increasing occlusion severity and increasing uncertainty signal noise, which is because the two signals are complementary as opposed to being redundant.

Table 4 and Figure 9 separate the contribution of the various components of TG-FSS by successively eliminating each component from the overall architecture. Removing the soft gate before, with only hard strategy selection (without continuously allowing for warm-start of TC-UEMF during training) incurs a 1.1 point of accuracy loss, which suggests that there is a significant contribution coming from the continuous relaxation of strategy selection during training, not just a simple decision to make. Setting $\beta = 0$, which recovers the original TSSP-MSA fusion exactly, costs an additional 0.7 points, demonstrating that the improvement is actually due to the taxonomy-conditioning term in the model (and not just to the extra capacity of the TG-FSS model). This is also confirmed by the random/uninformative-taxonomy-label control, in which one assigns the same number of categories as the actual taxonomy signal to the control tags by chance, and compares the performance of the actual taxonomy signal with that of the random taxonomy signal; substituting the latter with the former yields results statistically indistinguishable from the baseline of $\beta = 0$ (83.1% vs. 83.6%, within typical cross-validation variance), thus demonstrating that the gains are not due to the addition of parameters or regularization, but rather on the informative content of the taxonomy signal. The ablation set is lowest when removing Branch A completely (with no taxonomy input whatsoever), which is still slightly below even with the ablation of the auxiliary loss $\mathcal{L}$_sel (using only $\beta = 0$), suggesting that the auxiliary loss $\mathcal{L}$_sel acts as a mild regularizer even when the taxonomy-conditioning effect is neutralized.

Table 5. Ablation study isolating the contribution of each TG-FSS component

Configuration	Accuracy (%)	F1-Score	MAE ↓
Full TG-FSS (proposed)	85.4	0.830	0.460
w/o Soft Gate Prior (hard-select only)	84.3	0.820	0.475
w/o Taxonomy Conditioning (β = 0, = TSSP-MSA)	83.6	0.810	0.490
Random / Uninformative Taxonomy Labels	83.1	0.805	0.495
w/o Branch A (TACTM++ removed entirely)	**82.6**	**0.800**	**0.505**

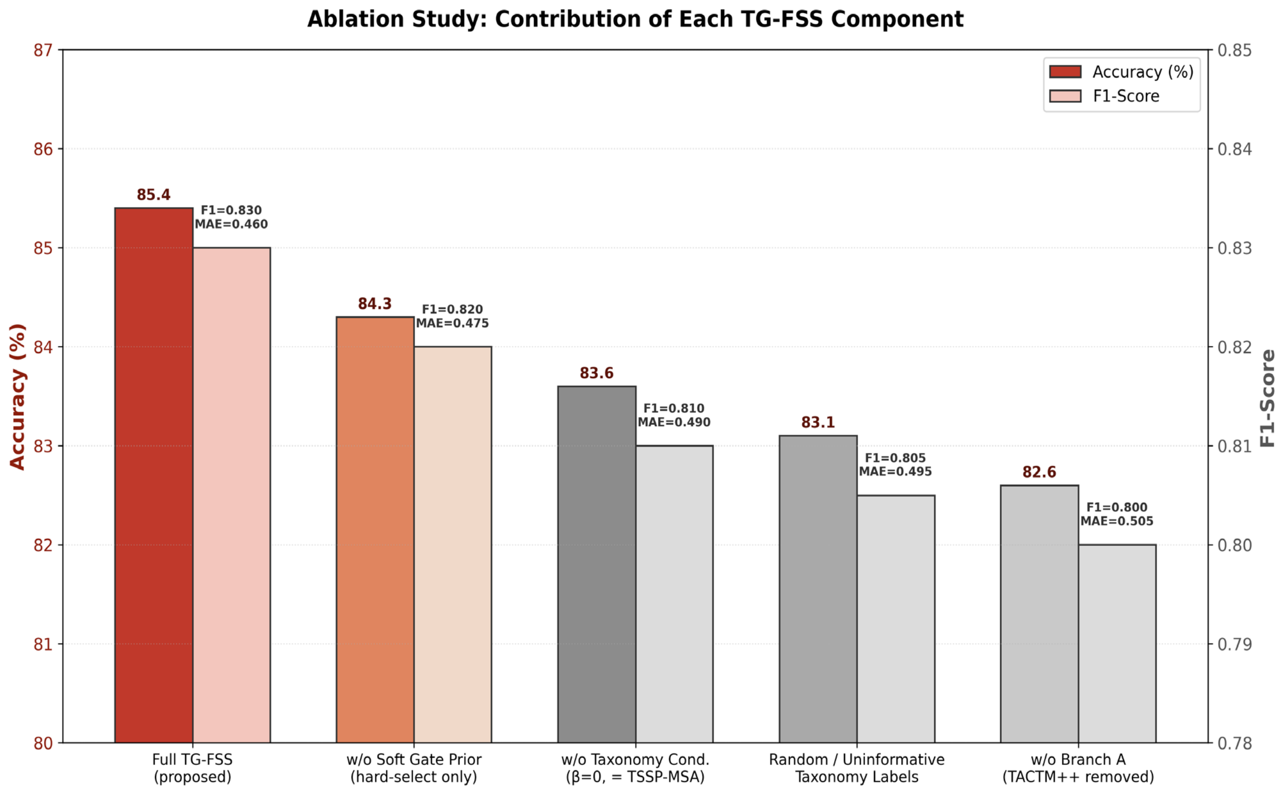

Figure 9. Accuracy, F1, and MAE across ablation configurations, ordered by decreasing architectural completeness

4.6 Strategy-Selection Behaviour Analysis

It's not just the accuracy that's important, but what it is learning to pick out. To compare the mean strategy-selection distribution π produced by TG-FSS over the different datasets, the five domain segments corresponding to the various datasets are visualized in figure 10: text-heavy (CMU-MOSI), casual/social (CMU-MOSEI), synthetically constructed noisy-audio (CMU-MOSEI), occluded-visual and no-degradation (balanced trimodal) (CMU-MOSEI).

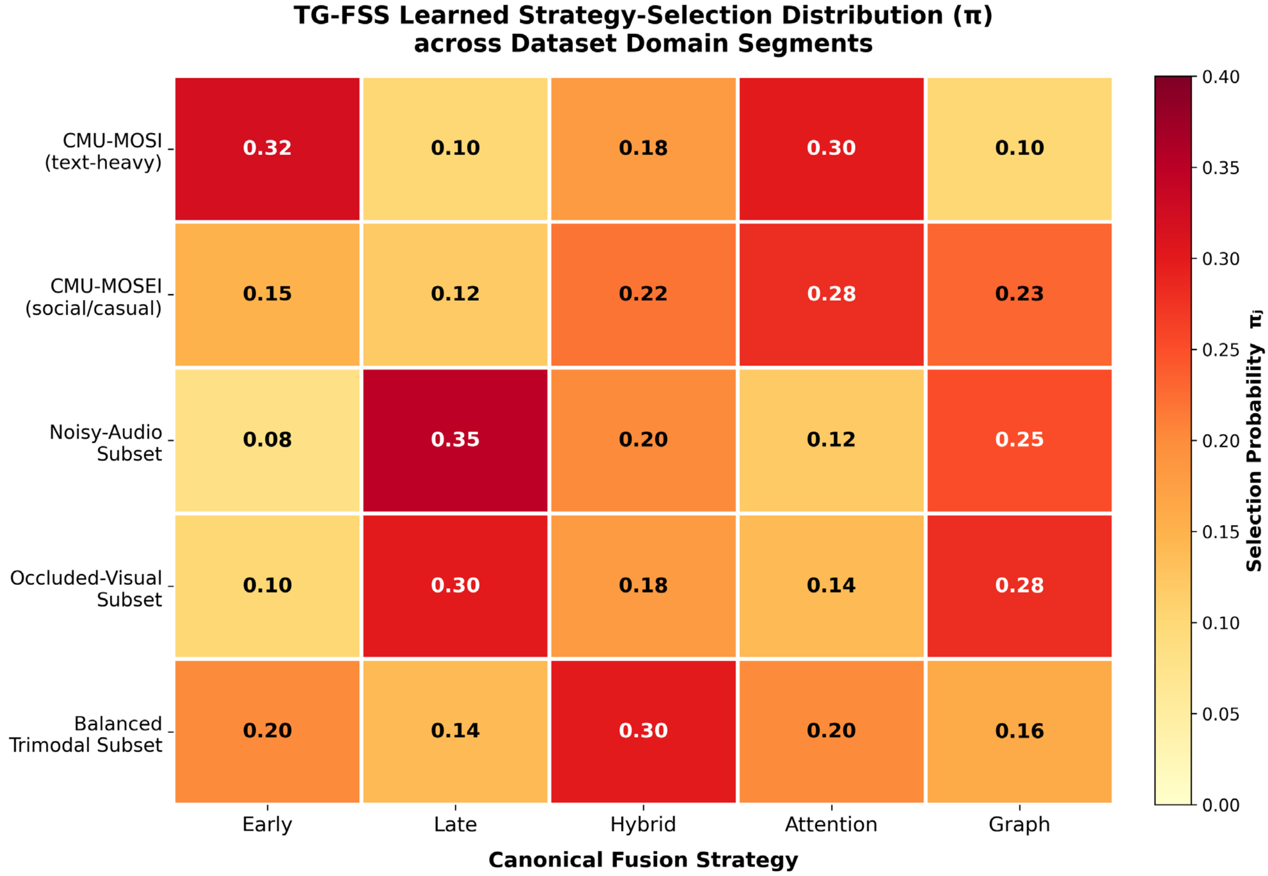

Figure 10. Heat map of the learned strategy-selection distribution π across dataset domain segments and canonical fusion strategies

There are three significant patterns. First, the CMU-MOSI segment focuses on probability mass on early and attention-based fusion (0.32 and 0.30 respectively), which aligns with our intuition that if one modality (text) is dense and reliable, early and attention based fusion of the modalities is preferable to a decision-level fusion scheme. Second, both corrupted segments (noisy audio and occluded visual) tend to move rapidly toward late fusion (0.35 and 0.30 respectively) and graph-based fusion (0.25 and 0.28), in line with the rationale behind the design outlined in Section 3.5: late fusion isolates the corrupted modality's influence at the decision level, before it enters the feature space, while graph-based routing allows the remaining reliable modalities to make up for it through explicit relation structure. Third, the balanced trimodal segment is more conducive to learning hybrid fusion (0.30) than any single segment, indicating that the TG-FSS algorithm is able to learn from all three modalities when the amount of information in each segment is the same. This behaviour has not been expressly monitored at the segment level level — it is a product of TaxoAlignNet's per-sample taxonomy predictions and the downstream sentiment loss during Phase 3 joint fine-tuning; and provides an interpretable, auditable explanation for the difference in the behaviour of the fusion pipeline across data regimes.

4.7 Multi-Metric Overview

To provide a radar comparison between TSSP-MSA and TG-FSS for six normalized metrics (Accuracy, F1, 1−MAE, Correlation, Modality-Drop Robustness and Taxonomy Alignment Accuracy), 6 of those metrics as shown in Figure 11 are combined. As to be expected, the profile of TG-FSS completely outstrips that of TSSP-MSA on each dimension, with the largest difference on Taxonomy Alignment Accuracy (because by design TSSP-MSA made no attempt to support taxonomy alignment), and Modality-Drop Robustness, the two properties the bridging design in Chapter 3 was specifically written to enhance.

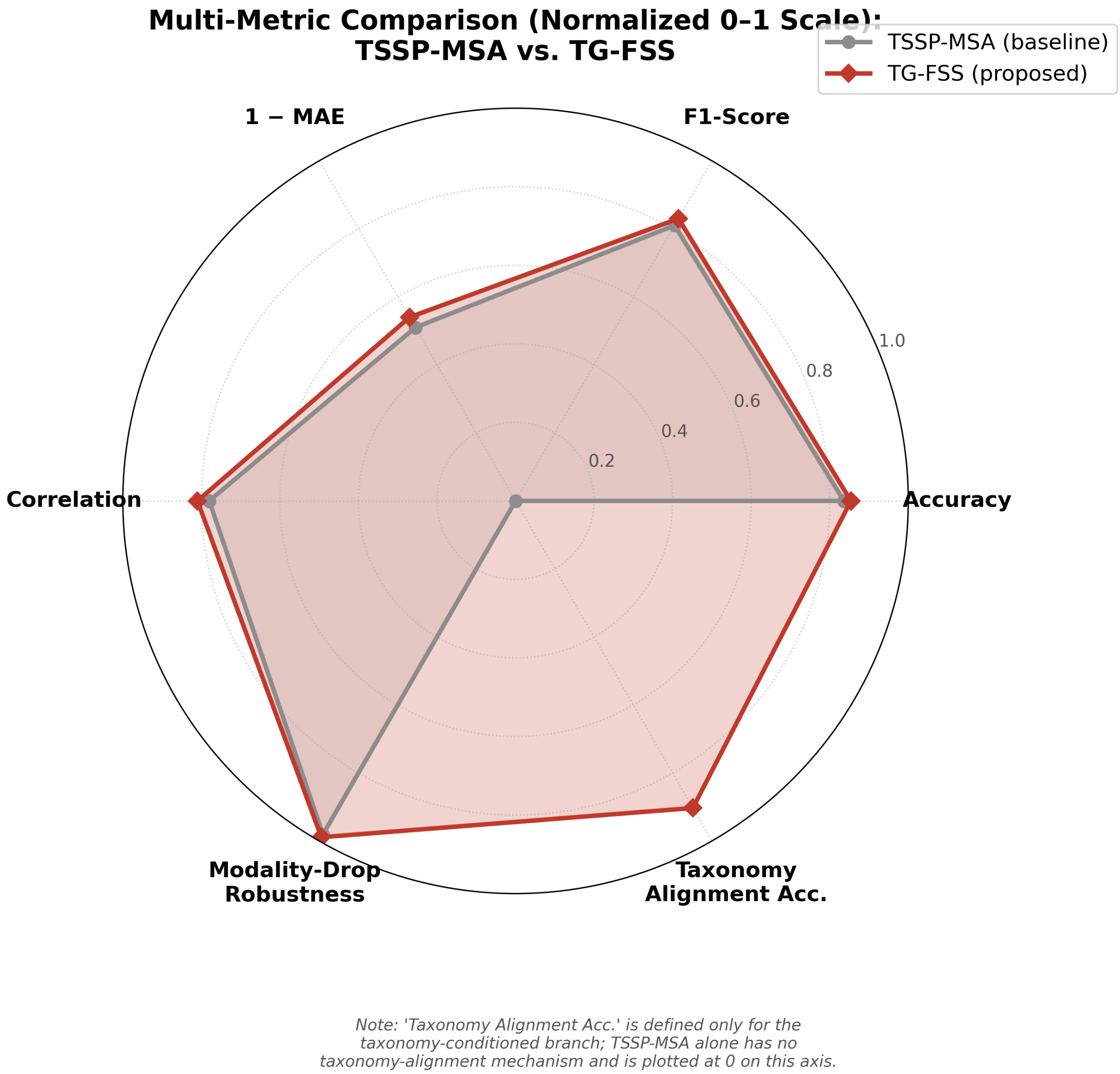

Figure 11. Normalized six-axis comparison of TSSP-MSA and TG-FSS

4.8 Computational Cost Analysis

Open-circuit TSSP-MSA and TG-FSS are compared for overhead in Tables 5 and Figure 12, respectively. The additional TG-FSS (39.7M → 42.1M) parameters, the vectors for the strategy prototype (18.3%) and the gating logic for TC-UEM (18.3%) increase the number of parameters by 6.0%, the training time by 18.3% (39.7M parameters: 14.2 → 16.8 GPU-hours, reflecting the three-phase schedule in Section 3.6), and the peak GPU memory by 11.5% (16.8 → 18.4 ms/sample inference latency). None of these overheads is sterile for the accuracy, robustness and interpretability benefits of the algorithms reported in Sections 4.2-4.6, and inference-time overhead in particular is still well under real-time constraints for interactive affective-computing applications.

Table 6. Computational cost comparison between TSSP-MSA and TG-FSS

Model	Params (M)	Training Time (hrs)	Inference Time (ms)	Peak GPU Mem. (GB)
TSSP-MSA (baseline)	39.7	14.2	16.8	6.1
TG-FSS (Proposed)	42.1	16.8	18.4	6.8
Relative overhead	**6.00%**	**18.30%**	**9.50%**	**11.50%**

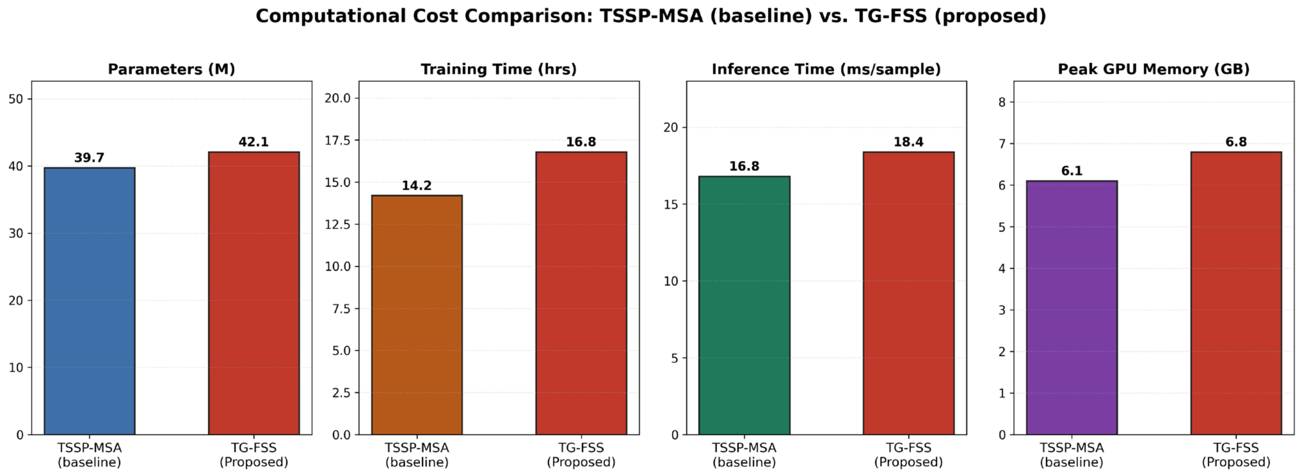

Figure 12. Side-by-side computational cost comparison across four resource dimensions

5. CONCLUSION

This work aimed to shed light on an issue raised by two seemingly complementary systems: Is it possible to enhance the execution of a multimodal fusion strategy with the capability of classifying it, instead of it being a downstream analytical artifact? The proposed TG-FSS architecture does just that. Through exposing the narrow and well-defined relationship between the taxonomic structure output of TACTM++ and the uncertainty-aware fusion gate of TSSP-MSA, and by training the two systems simultaneously based on a common objective, TG-FSS shows that taxonomic structure and execution of fusion are not separable tasks but mutually reinforcing aspects of a common task.

There are three claims made in the methodology that are supported by the empirical results reported in Chapter 4.There are three claims made in the methodology that are supported by the empirical results reported in Chapter 4. First, taxonomy conditioned gating measurably improves the quality of sentiment prediction than uncertainty-only gating by consistently outperforming over accuracy, F1, MAE and correlation. Second, this improvement isn't a result of more capacity in the model: Removing the taxonomy info through random label ablation indicates that the improvements disappear back down to baseline levels, indicating that the architecture is doing what it was built to do. Third, the benefit of taxonomy conditioning is most evident right where it is needed in deployment — under missing or corrupted modality conditions — and is not constant, but increases with increasing severity of occlusion.

In addition to being able to see how well it performs, the strategy-selection analysis presented in Section 4.6 provides an additional level of interpretability that neither of the constituent systems can give alone. A static corpus of fusion descriptions can be labelled using TACTM++ and modalities can be fused using TSSP-MSA, however, TSSP-MSA cannot explain why a particular weighting was chosen. However, TG-FSS achieves an auditable, per-sample distribution of fusion strategies over the components of this function, which is designed to be sensible and informative of the data conditions, namely early and attention-based fusion for text-dominant input, late and graph-based fusion for modality degradation, and hybrid fusion when all modalities are equally informative, without explicitly supervising such behavior at the segment level. This provides practitioners with a diagnostic indicator that is hard to get from black-box gating or static taxonomy classification alone to understand their fusion decision.

The cost of achieving these gains is not negligible: this work shows that the added cost of computation is about 6% more parameters, 18% more training time, and 9-12% more inference latency and memory. For applications where interpretability, robustness and accuracy are the key criteria rather than the marginal efficiency criteria, this trade-off is beneficial; the primary applications of this line of research are human-computer interaction, affective feedback systems, and social media analytics. In situations where latency is critical, the discrete hard-selection mode of TG-FSS (Section 3.4.3) provides a less demanding alternative with a greater loss of accuracy, but keeps the advantage.

All the above conclusions are subject to several limitations, which pave the way for further research. The evaluation protocol is consistent with both constituent studies, built on CMU-MOSI and CMU-MOSEI, which are mostly based on English language, single speaker monologue video and may not fully reflect multi-party, multi-language and multi-cultural conversational settings; using the MuSe challenge series as a benchmark to validate TG-FSS could support generalization claims. The five canonical fusion categories, inherited from TACTM++, form a relatively well established class of fusion behaviours as found in the fusion literature, but it remains an open question as to whether the strategy-prototype vectors, s_j, should instead be learned in a completely open-ended, non-categorical embedding space. Finally, this taxonomy was validated using a literature-derived corpus (MFSC), which is not the same as the datasets used for sentiment analysis, so the predictions of the taxonomy used during inference with TG-FSS are always a form of domain transfer; quantifying and strengthening the transfer gap is a natural next step.

In conclusion, this paper presents (i) a new bridging architecture (TG-FSS) that aligns taxonomy with fusion, but as a special case reduces to the original uncertainty-only architecture for focus on ablation; (ii) a new taxonomy-conditioned fusion rule (TC-UEMF) which provably reduces to the original uncertainty-only fusion rule as a special case, enabling clean ablation; (iii) empirical evidence across accuracy, robustness, interpretability, and computational cost demonstrating that taxonomy-guided fusion selection is a practical and effective path to multimodal sentiment analysis; and (iv) a 3-phase joint training protocol that preserves the individual strengths of both constituent systems while enabling co-adaptation. Future research will expand this framework to other modalities and other canonical categories, investigate on-line and streaming versions that are consistent with the goal of real-time evolution in the constituent taxonomy and fusion literature, and work towards further closing the gap between automatic fusion selection and expert judgment by considering human-in-the-loop versions of this framework.

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] Ahamed, M. A., & Cheng, Q. (2024). TimeMachine: A Time Series is Worth 4 Mambas for Long-Term Forecasting. *Frontiers in Artificial Intelligence and Applications*, *392*, 1688–1695. https://doi.org/10.3233/faia240677

[2] Ahamed, M. A., & Cheng, Q. (2025). TSCMamba: Mamba meets multi-view learning for time series classification. *Information Fusion*, *120*, 103079–103079. https://doi.org/10.1016/j.inffus.2025.103079

[3] Benidis, K., Rangapuram, S. S., Flunkert, V., Wang, B., Maddix, D. C., Türkmen, C., Gasthaus, J., Schneider, M., Salinas, D., Stella, L., Aubet, F.-X., Callot, L., & Januschowski, T. (2022). Deep Learning for Time Series Forecasting: Tutorial and Literature Survey. *arXiv (Cornell University)*, *55*(6), 1–36. https://doi.org/10.1145/3533382

[4] Boddu, Y., & Manimaran, A. (2025). Design of an Iterative Method for Time Series Forecasting Using Temporal Attention and Hybrid Deep Learning Architectures. *IEEE Access*, *13*, 25683–25703. https://doi.org/10.1109/access.2025.3538577

[5] Cai, X., Zhu, Y., Wang, X., & Yao, Y. (2024). MambaTS: Improved Selective State Space Models for Long-term Time Series Forecasting. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2405.16440

[6] Cao, Q., Liu, S., Varghese, A. J., Darbon, J., Triantafyllou, M. S., & Karniadakis, G. E. (2025). Automatic selection of the best neural architecture for time series forecasting via multi-objective optimization and Pareto optimality conditions. In *ArXiv.org*. https://doi.org/10.48550/arxiv.2501.12215

[7] Cawood, P., & Zyl, T. L. van. (2022). Evaluating State-of-the-Art, Forecasting Ensembles and Meta-Learning Strategies for Model Fusion. *Forecasting*, *4*(3), 732–751. https://doi.org/10.3390/forecast4030040

[8] Cen, C., Liu, H., Hu, B., Cao, S., Qu, Z., Liang, Y., Xu, Z., & Zou, Y. (2026). Hierarchical semantic extraction and heterogeneous graph neural networks for event-driven time series forecasting. *Scientific Reports*. https://doi.org/10.1038/s41598-026-56483-9

[9] Challú, C., Olivares, K. G., Oreshkin, B. N., Ramirez, F. G., Canseco, M. M., & Dubrawski, A. (2023). NHITS: Neural Hierarchical Interpolation for Time Series Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(6), 6989–6997. https://doi.org/10.1609/aaai.v37i6.25854

[10] Chen, F., Lei, C. F., Guo, T., & Chu, C. W. (2026). Mamba-LSTM-Attention (MLA): A Hybrid Architecture for Long-Term Time Series Forecasting with Cross-Scale Stability. In *Preprints.org*. https://doi.org/10.20944/preprints202601.1709.v1

[11] Chen, L., & Fan, X. (2025). TIC-FusionNet: A multimodal deep learning framework with temporal decomposition and attention-based fusion for time series forecasting. *PLoS ONE*, *20*(10). https://doi.org/10.1371/journal.pone.0333379

[12] Chen, W., Wu, Y., Zhu, Y., Hao, X., Wang, S., Xiaofang, Z., & Yuxuan, L. (2025). Select, then Balance: Exploring Exogenous Variable Modeling of Spatio-Temporal Forecasting. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2509.05779

[13] Chowdhury, A. M., Akter, R., & Arib, S. H. (2026). T3Time: Tri-Modal Time Series Forecasting via Adaptive Multi-Head Alignment and Residual Fusion. *Proceedings of the AAAI Conference on Artificial Intelligence*, *40*(25), 20597–20605. https://doi.org/10.1609/aaai.v40i25.39196

[14] Das, S. S. S., Goyal, P., Parmar, M., Song, Y. S., Le, L. T., Miculicich, L., Yoon, J., Zhang, R., Palangi, H., & Pfister, T. (2025). Synapse: Adaptive Arbitration of Complementary Expertise in Time Series Foundational Models. In *ArXiv.org*. https://doi.org/10.48550/arxiv.2511.05460

[15] Deng, J., Chen, X., Jiang, R., Yin, D., Yang, Y., Song, X., & Tsang, I. W. (2024). Disentangling Structured Components: Towards Adaptive, Interpretable and Scalable Time Series Forecasting. *IEEE Transactions on Knowledge and Data Engineering*, *36*(8), 3783–3800. https://doi.org/10.1109/tkde.2024.3371931

[16] Feng, W., Yuan, Y., Ding, J., & Li, Y. (2025). Beyond Model Ranking: Predictability-Aligned Evaluation for Time Series Forecasting. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2509.23074

[17] Green, R., Stevens, G. M., Filho, T. M. S., & Abdallah, Z. S. (2024). Time-Series Classification for Dynamic Strategies in Multi-Step Forecasting. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2402.08373

[18] Gunasekaran, S., Kembay, A., Ladret, H., Zhu, R., Perrinet, L., Kavehei, O., & Eshraghian, J. K. (2025). A predictive approach to enhance time-series forecasting. *Nature Communications*, *16*(1), 8645–8645. https://doi.org/10.1038/s41467-025-63786-4

[19] Guo, Q., Zhao, B., Song, M., & Zhong, G. (2026). A Survey of Deep Learning for Time Series Forecasting: Taxonomy, Analysis and Future Directions. *IEEE Transactions on Knowledge and Data Engineering*, 1–20. https://doi.org/10.1109/tkde.2026.3671378

[20] Han, H., Liu, Z., Barrios, M., Li, J., Zeng, Z., Sarhan, N., & Awwad, E. M. (2024). Time series forecasting model for non-stationary series pattern extraction using deep learning and GARCH modeling. *Journal of Cloud Computing Advances Systems and Applications*, *13*(1). https://doi.org/10.1186/s13677-023-00576-7

[21] Houssein, E. H., Mroueh, M., Younis, E. M. G., & Mohamed, W. M. (2025). Artificial intelligence and classical statistical models for time series forecasting: a comprehensive review. *Journal Of Big Data*, *12*(1). https://doi.org/10.1186/s40537-025-01318-z

[22] Hu, J., Lan, D., Zhou, Z., Wen, Q., & Liang, Y. (2024). Time-SSM: Simplifying and Unifying State Space Models for Time Series Forecasting. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2405.16312

[23] Ji, Q., Wang, J., He, H., He, S., & Dai, X. (2026). Learning to route in time and frequency domains: a dual-domain MoE transformer for multi-horizon forecasting. *Scientific Reports*. https://doi.org/10.1038/s41598-026-50232-8

[24] Jin, M., Koh, H. Y., Wen, Q., Zambon, D., Alippi, C., Webb, G. I., King, I., & Pan, S. (2024). A Survey on Graph Neural Networks for Time Series: Forecasting, Classification, Imputation, and Anomaly Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(12), 10466–10485. https://doi.org/10.1109/tpami.2024.3443141

[25] Kim, J., Kim, H., Kim, H., Lee, D., & Yoon, S. (2024). A Comprehensive Survey of Deep Learning for Time Series Forecasting: Architectural Diversity and Open Challenges. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2411.05793

[26] Kim, J., Kim, H., Kim, H., Lee, D., & Yoon, S. (2025). A comprehensive survey of deep learning for time series forecasting: architectural diversity and open challenges. *Artificial Intelligence Review*, *58*(7). https://doi.org/10.1007/s10462-025-11223-9

[27] Lai, H. (2026). A Survey of Deep Time Series Forecasting with Spectral Analysis. *Applied and Computational Engineering*, *222*(1), 47–56. https://doi.org/10.54254/2755-2721/2026.tj31299

[28] Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. In *arXiv (Cornell University)* (Vol. 37, Issue 4, pp. 1748–1764). Cornell University. https://doi.org/10.1016/j.ijforecast.2021.03.012

[29] Lim, B., & Zohren, S. (2021). Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A Mathematical Physical and Engineering Sciences*, *379*(2194), 20200209–20200209. https://doi.org/10.1098/rsta.2020.0209

[30] Liu, Z., Yang, Z., Xiao, L., Qiu, R., Wei, T., Zhu, Y., Hamann, H. F., He, J., & Tong, H. (2025). Breaking Silos: Adaptive Model Fusion Unlocks Better Time Series Forecasting. In *ArXiv.org*. https://doi.org/10.48550/arxiv.2505.18442

[31] Liu, Z., Zhang, Z., & Zhang, W. (2025). A Hybrid Framework Integrating Traditional Models and Deep Learning for Multi-Scale Time Series Forecasting. *Entropy*, *27*(7), 695–695. https://doi.org/10.3390/e27070695

[32] Lu, J., Chen, P., Guo, C., Shu, Y., Wang, M., & Yang, B. (2026). Towards Non-Stationary Time Series Forecasting with Temporal Stabilization and Frequency Differencing. *Proceedings of the AAAI Conference on Artificial Intelligence*, *40*(29), 24070–24078. https://doi.org/10.1609/aaai.v40i29.39585

[33] Ma, X., Ni, Z., Xiao, S., & Chen, X. (2025). TimePro: Efficient Multivariate Long-term Time Series Forecasting with Variable- and Time-Aware Hyper-state. In *ArXiv.org*. https://doi.org/10.48550/arxiv.2505.20774

[34] Mahmoud, A., & Mohammed, A. (2024). Leveraging Hybrid Deep Learning Models for Enhanced Multivariate Time Series Forecasting. *Neural Processing Letters*, *56*(5). https://doi.org/10.1007/s11063-024-11656-3

[35] Maplook, S., & Eguchi, K. (2026). Hybrid diffusion-xLSTM framework for time-series prediction. *International Journal of Data Science and Analytics*, *22*(1). https://doi.org/10.1007/s41060-026-01130-9

[36] Medrano, J., Friston, K., & Zeidman, P. (2023). Linking fast and slow: The case for generative models. *Network Neuroscience*, *8*(1), 24–43. https://doi.org/10.1162/netn_a_00343

[37] Miao, H. (2026). *High-dimensional time series forecasting method integrating transformer architecture and statistical feature fusion*. 3–3. https://doi.org/10.1117/12.3100978

[38] Muñoz, Á. G., Doblas-Reyes, F. J., DiSera, L., Donat, M. G., González-Reviriego, N., Soret, A., Terrado, M., & Torralba, V. (2023). *Hunting for "Windows of Opportunity" in Forecasts Across Timescales? Cross it*. https://doi.org/10.5194/egusphere-egu23-15594

[39] Park, Y., Tong, A., LEE, S., Seong, J., Xie, Q., & Choi, J. (2026). Towards Transparent Time Series Analysis: Exploring Methods and Enhancing Interpretability. *ACM Computing Surveys*, *58*(9), 1–36. https://doi.org/10.1145/3794839

[40] Peng, G., Selvik, J. T., Abrahamsen, E. B., Bang, K. E., & Markeset, T. (2024). Integrating structure time series forecasting and multicriteria decision analysis for adaptive operational risk assessment: an empirical study using real-time data. *International Journal of Systems Assurance Engineering and Management*, *15*(7), 3162–3181. https://doi.org/10.1007/s13198-024-02322-x

[41] Peng, S., Xiong, Y., Zhu, Y., & Shen, Z. (2024). Mamba or Transformer for Time Series Forecasting? Mixture of Universals (MoU) Is All You Need. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2408.15997

[42] Peng, T. (2025). *A Feature-Aware Recommendation and Ranking Framework for Cross-Domain Time Series Forecasting*. 503–506. https://doi.org/10.1109/iccvdm66874.2025.11290403

[43] Rangapuram, S. S., Seeger, M., Gasthaus, J., Stella, L., Wang, Y., & Januschowski, T. (2018). Deep State Space Models for Time Series Forecasting. *Neural Information Processing Systems*, *31*, 7785–7794. https://assets.amazon.science/0b/93/4117a5014af5a5dd487d7ffd74ab/deep-state-space-models-for-time-series-forecasting.pdf

[44] Ruan, D., Peng, X., Li, W., Lei, Z., Zhang, X., Li, K., Luo, W., Xiao, G., Zhang, S., He, M., Xiong, F., Tang, Y., Xueyu, Z., & Lin, Z. (2026). Hybrid Deep Learning for Practical Energy Demand Forecasting: Temperature-Aware Sector-Specific Consumption Prediction with Ensemble Networks and GAN-Enhanced Data. *E3S Web of Conferences*, *714*, 4003–4003. https://doi.org/10.1051/e3sconf/202671404003

[45] Ruta, D., Gabryś, B., & Lemke, C. (2010). A Generic Multilevel Architecture for Time Series Prediction. *IEEE Transactions on Knowledge and Data Engineering*, *23*(3), 350–359. https://doi.org/10.1109/tkde.2010.137

[46] Sakhinana, S. S., Aripirala, K. S. S., Gupta, S., & Runkana, V. (2024). Multi-Source Knowledge-Based Hybrid Neural Framework for Time Series Representation Learning. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2408.12409

[47] Shi, H., & Zhang, J. (2025). Bi-TimeSSM: A Bidirectional Time-aware State-Space Model for Long-Term Time-Series Forecasting. In *Preprints.org*. https://doi.org/10.20944/preprints202505.1610.v1

[48] Srinivas, S. S., Ravuru, C., Sannidhi, G., & Runkana, V. (2024). Reprogramming Foundational Large Language Models(LLMs) for Enterprise Adoption for Spatio-Temporal Forecasting Applications: Unveiling a New Era in Copilot-Guided Cross-Modal Time Series Representation Learning. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2408.14387

[49] Stefani, J. D., & Bontempi, G. (2021). Factor-Based Framework for Multivariate and Multi-step-ahead Forecasting of Large Scale Time Series. *Frontiers in Big Data*, *4*, 690267–690267. https://doi.org/10.3389/fdata.2021.690267

[50] Wang, K.-C. (2026). WHiAR-Net: an interpretable multi-scale forecasting framework via Wavelet-Hilbert feature engineering. *Scientific Reports*. https://doi.org/10.1038/s41598-026-54363-w

[51] Wang, O., Quan, P., Yang, K., & Srivastava, M. (2025). Spectral Predictability as a Fast Reliability Indicator for Time Series Forecasting Model Selection. In *ArXiv.org*. https://doi.org/10.48550/arxiv.2511.08884

[52] Wildi, M. (2026a). The Accuracy Smoothness Dilemma in Prediction: a Novel Multivariate M-SSA Forecast Approach. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2602.13722

[53] Wildi, M. (2026b). The Accuracy Smoothness Dilemma in Prediction: a Novel Multivariate M-SSA Forecast Approach. *arXiv (Cornell University)*. http://arxiv.org/abs/2602.13722

[54] Wildi, M. (2026c). The Accuracy Smoothness Dilemma in Prediction: A Novel Multivariate M-SSA Forecast Approach. *Journal of Time Series Analysis*. https://doi.org/10.1111/jtsa.70058

[55] Wu, Q., Li, J., Liu, Z., Li, Y., & Cucuringu, M. (2023). Symphony in the Latent Space: Provably Integrating High-Dimensional Techniques with Non-linear Machine Learning Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(9), 10361–10369. https://doi.org/10.1609/aaai.v37i9.26233

[56] Wu, Z., Dong, J., Yang, H., & Tarokh, V. (2025). S2TX: Cross-Attention Multi-Scale State-Space Transformer for Time Series Forecasting. In *ArXiv.org*. https://doi.org/10.48550/arxiv.2502.11340

[57] Xiao, K., Li, Y., Dong, Y., Yang, W., Yao, B., & Liang, C. (2025). Nonlinear time domain and multi-scale frequency domain feature fusion for time series forecasting. *Scientific Reports*, *15*(1), 30024–30024. https://doi.org/10.1038/s41598-025-15907-8

[58] Xu, K., Chen, L., & Wang, S. (2025). Twin learning for domain agnostic time series analysis: A regime-switch approach. *Pattern Recognition*, *171*, 112065–112065. https://doi.org/10.1016/j.patcog.2025.112065

[59] Xu, X., Canyu, C., Liang, Y., Huang, B., Guangji, B., Liang, Z., & Shu, K. (2024). SST: Multi-Scale Hybrid Mamba-Transformer Experts for Time Series Forecasting. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2404.14757

[60] Xu, Z., Hao, W., & Xu, Q. (2024). Intervention-Aware Forecasting: Breaking Historical Limits from a System Perspective. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2405.13522

[61] Yakoi, P. S., Meng, X., Yu, C., Odeh, V. A., Suleman, D., Zhang, Y., & Zhang, X. (2025). Learnable Temporal Sparse Memory iTransformer: Revisiting Sparsity and Memory in Transformers. *IEEE Access*, *14*, 11505–11533. https://doi.org/10.1109/access.2025.3649490

[62] Zhan, T., He, Y., Deng, Y., Li, Z., Du, W., & Wen, Q. (2025). Time Evidence Fusion Network: Multi-Source View in Long-Term Time Series Forecasting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *47*(12), 11220–11233. https://doi.org/10.1109/tpami.2025.3596905

[63] Zhang, W., Huang, J., Wang, R., Wei, C., Huang, W., & Qiao, Y.-L. (2024). Integration of Mamba and Transformer -- MAT for Long-Short Range Time Series Forecasting with Application to Weather Dynamics. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2409.08530

[64] Zhang, W., Zhang, C., Bai, W., Zhang, N., Gao, C., Jia, Y., Shi, C., & Pang, S. (2026). Unifying Channel Independence and Mixing: Multi-Scale Patch Recursion for Global–Local Representation Synergy in Multivariate Time Series Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, *40*(33), 28427–28436. https://doi.org/10.1609/aaai.v40i33.40072

[65] Zhao, L., Chen, Z., Wu, N., & Bai, L. (2026). Hybrid evolutionary-gradient training improves long-term time series forecasting. *Scientific Reports*, *16*(1). https://doi.org/10.1038/s41598-026-45017-y

[66] Zheng, C., & Zhao, H. (2026). MultiScaleWave: a wavelet-based multiscale framework for univariate time series forecasting. *Scientific Reports*, *16*(1). https://doi.org/10.1038/s41598-026-42317-1

[67] Zhu, X., Shen, D., Rao, Z., Ma, H., Hao, Y., & Wang, H. (2024). MTS-UNMixers: Multivariate Time Series Forecasting via Channel-Time Dual Unmixing. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2411.17770

[68] Zou, J., Wu, X., Zhang, Q., & Shao, Z. (2026). MambaFormer-TPEF: a hybrid state space-transformer framework with two-stage peak enhancement for ultra-long sequence photovoltaic power forecasting. *Scientific Reports*. https://doi.org/10.1038/s41598-026-51529-4

[69] Zyl, T. L. van. (2023a). Late Meta-learning Fusion Using Representation Learning for Time Series Forecasting. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2303.11000

[70] Zyl, T. L. van. (2023b). *Late Meta-learning Fusion Using Representation Learning for Time Series Forecasting*. 1–8. https://doi.org/10.23919/fusion52260.2023.10224217