MULTIMODAL FUSION ON THE BOOLEAN LATTICE: FROM LOCAL CONTEXT DEPENDENCE TO GLOBAL HIGHER-ORDER INTERACTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

In multimodal fusion, a modality's marginal contribution can depend on the combination of modalities already present, which defines its context. Viewing fusion as a vector-valued set function over modality subsets shows that this context dependence is equivalent to higher-order interactions of order two or above. We therefore develop a formulation of multimodal fusion on the Boolean lattice of modality subsets, capturing global higher-order interactions by structuring local context dependence. Building on this formulation, we introduce LATFUSE, which constructs fused representations for each subset of modalities in a bottom-up manner on the Boolean lattice, using single-modality extensions as reusable units for fusion. Across four benchmarks comprising three or four modalities and spanning multiomics classification, sentiment analysis, and semantic segmentation, LATFUSE outperforms the strongest baselines on all task metrics. By constructing every subset state, LATFUSE can infer directly on sublattices and is more robust than the baselines under most missing-modality conditions. On synthetic tasks with known interaction structures, LATFUSE learns interactions that largely match the true interaction orders, supporting analysis of higher-order interactions.

1 INTRODUCTION

Multimodal fusion often requires accounting for non-additive higher-order interactions, which arise only from two or more modalities together and cannot be explained by each modality's independent effect (Wenderoth et al., 2025). A modality's marginal contribution to fusion depends on the other modalities present, which form its context (Sundararajan et al., 2020; Amit et al., 2025). For example, with modalities $\{A, B, C\}$, adding C to $\{A\}$ and to $\{A, B\}$ may change the fused representation differently. Such context-dependent interaction patterns typically vary across tasks and data distributions (Madaan et al., 2025; Yang et al., 2025a), so they should be represented and learned by the fusion models themselves. This suggests that modelling higher-order interactions requires examining how a modality's marginal contribution varies with its context.

Many models can implicitly express higher-order interactions, but these abilities do not always manifest as the interactions required by the tasks (Hessel & Lee, 2020; Yang et al., 2026). Previous research has explicitly modelled joint representations of multiple modalities, for example, using tensor product expansion (Zadeh et al., 2017; Liu et al., 2018) or residual prediction over modality subsets (Wörtwein et al., 2022). Even with explicit modelling, the interaction types captured by a fusion architecture need not match those required by a given task (Liang et al., 2023). Adaptive fusion adjusts each modality's influence with sample-dependent weights or gates (Zhang et al., 2023c), but rarely addresses interactions among combinations of modalities. These existing approaches tend to treat higher-order interactions and context dependence separately: higher-order interactions are characterised at the level of the global fused representation, whereas context dependence concerns how an individual modality behaves under different modality combinations. Bridging these two perspectives requires a common object that characterises how a modality's contribution varies with context and how such variation gives rise to higher-order interactions.

We establish this connection by viewing multimodal fusion as a vector-valued set function that maps each subset of modalities to a fused representation in a common vector space. Combined with

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

discrete-difference characterisations of interactions (Grabisch & Labreuche, 2007), this formulation yields an equivalence that existing research on multimodal fusion has not fully exploited: *the fusion function exhibits higher-order interactions if and only if the marginal contribution of some modality to the fused representation varies with the modalities already present*. This equivalence motivates a principle: we can directly parameterise context-conditioned single-modality extensions with explicit semantics and compose global higher-order interactions from these local extensions.

This principle admits a natural realisation on the Boolean lattice of modality subsets. Ordered by inclusion, these subsets form a lattice in which each covering edge $\mathcal{V} \xrightarrow{+m} \mathcal{V} \cup \{m\}$ represents the addition of modality m to context $\mathcal{V}$. The difference between the representations at the two endpoints corresponds to the marginal contribution of m in that context. We instantiate this structure in LATFUSE, which constructs fused representations for every non-empty subset in a bottom-up manner along the covering edges of the Boolean lattice (Figure 1b). A shared single-modality extension serves as the basic unit of fusion. This construction also handles missing modalities, since each subset representation depends only on the modalities it contains; when some modalities are absent, LATFUSE operates on the Boolean sublattice induced by the observed modalities.

We evaluate LATFUSE on four benchmarks with three or four modalities, spanning multiomics classification, sentiment analysis, and semantic segmentation. (*a*) When all modalities are available, LATFUSE outperforms the compared baselines on all task metrics (§5.2). Compared to the strongest baseline on each metric, LATFUSE achieves gains of 1.69 macro F_1 points on TCGA-BRCA and 2.88 ACC_7 points on CMU-MOSI. (*b*) Under most conditions with a missing modality, LATFUSE degrades less than the baselines, showing greater robustness at inference (§5.4). (*c*) On synthetic datasets with known interaction structures, fANOVA (Sobol', 2001) shows that LATFUSE largely recovers the true interaction orders (§5.5).

This paper makes the following contributions:

- We formalise multimodal fusion as a vector-valued set function and, through the equivalence between higher-order interactions and context-dependent marginal contributions, provide explicit local semantics for modelling and composing higher-order multimodal interactions (§2).
- We introduce LATFUSE, a fusion framework on the Boolean lattice that realises this formulation by modelling context-conditioned single-modality extensions and composing them into representations of arbitrary modality subsets. The same model supports different observed modality combinations, including inference with missing modalities (§3).
- We evaluate LATFUSE across multiple tasks and from three complementary perspectives: predictive performance, robustness to missing modalities, and recovery of known interaction structures. Across these settings, LATFUSE improves predictive performance and subset inference while largely recovering the underlying interaction structures (§5).

2 MULTIMODAL FUSION AS A SET FUNCTION

We formalise multimodal fusion as a vector-valued set function over modality subsets. Locally, covering-edge differences represent context-specific marginal contributions; globally, Möbius inversion yields interaction components. These two views are equivalent as characterisations of the same set function and therefore provide a unified description of multimodal interaction structure.

Subset states and the Boolean lattice. For a fixed sample, let $[M] \coloneqq \{1, \dots, M\}$ index the modalities and $\boldsymbol{x} = (\boldsymbol{x}_m)_{m \in [M]}$ collect their representations. The set function $H_{\boldsymbol{x}} : 2^{[M]} \to \mathbb{R}^{d_h \times N}$ maps each modality subset $\mathcal{U}$ to a fused representation $H_{\boldsymbol{x}}(\mathcal{U}) \coloneqq \boldsymbol{h}_{\mathcal{U}}$ with N tokens of width d_h. We call $\boldsymbol{h}_{\mathcal{U}}$ the *subset state* of $\mathcal{U}$ and fix the boundary state $\boldsymbol{h}_{\varnothing} \coloneqq \mathbf{0}$. Ordered by inclusion, the subsets form a Boolean lattice. A subset $\mathcal{U}$ has order $|\mathcal{U}|$, and $\mathcal{G}_r \coloneqq \{\mathcal{U} \subseteq [M] : |\mathcal{U}| = r\}$ collects the subsets of order r for $r = 1, \dots, M$.

Covering edges and marginal contributions. A *covering edge* $e : \mathcal{V} \xrightarrow{+m} \mathcal{U}$ adds modality $m \notin \mathcal{V}$ to context $\mathcal{V}$, giving $\mathcal{U} = \mathcal{V} \cup \{m\}$. Let $\mathcal{E}$ denote the set of covering edges. For any set function F with values in a real vector space, define the *forward difference* along e as

$$\Delta_e F \coloneqq (\Delta_m F)(\mathcal{V}) \coloneqq F(\mathcal{U}) - F(\mathcal{V}). \tag{1}$$

The difference $\Delta_e H_{\boldsymbol{x}} = \boldsymbol{h}_{\mathcal{U}} - \boldsymbol{h}_{\mathcal{V}}$ is the *marginal contribution* of modality m in context $\mathcal{V}$.

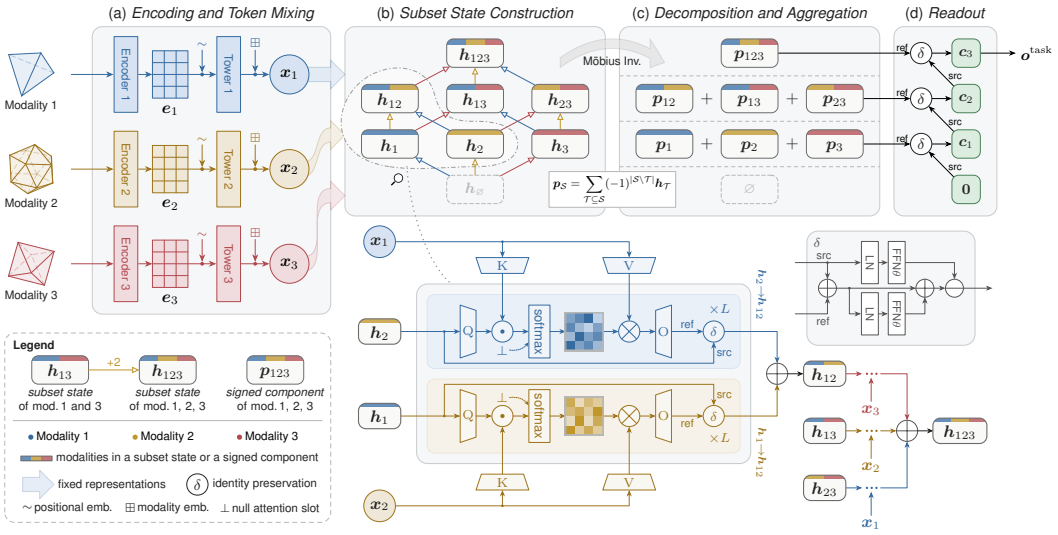


Figure 1: Overview of LATFUSE for three modalities. (a) Encoders and token-mixing towers produce modality representations $\boldsymbol{x}$ (§3.1). (b) Context-conditioned single-modality extensions construct subset states $\boldsymbol{h}_{(\cdot)}$ on the Boolean lattice (§3.3), where the $\wp$ expands the edge refinement operators (§3.2). (c) Möbius inversion yields signed interaction components $\boldsymbol{p}_{(\cdot)}$, aggregated by order (§3.4). (d) Cross-order residual fusion combines the order representations for task prediction (§§3.4 and 3.5).

Interaction components. While covering-edge differences describe interactions locally, Möbius inversion provides a complementary global decomposition of the set function. For each non-empty $\mathcal{S} \subseteq [M]$, Möbius inversion defines a *signed interaction component* of order $|\mathcal{S}|$ (Rota, 1964):

$$\boldsymbol{p}_{\mathcal{S}} := \sum_{\mathcal{T} \subseteq \mathcal{S}} (-1)^{|\mathcal{S} \setminus \mathcal{T}|} \boldsymbol{h}_{\mathcal{T}} \in \mathbb{R}^{d_h \times N}. \tag{2}$$

We call non-zero components with $|\mathcal{S}| \geqslant 2$ *higher-order interactions* of $H_{\boldsymbol{x}}$.

Equivalence to context dependence. Along $e : \mathcal{V} \xrightarrow{+m} \mathcal{U}$, reconstruction gives $\Delta_e H_{\boldsymbol{x}} = \sum_{\mathcal{T} \subseteq \mathcal{V}} \boldsymbol{p}_{\mathcal{T} \cup \{m\}}$ (Grabisch & Labreuche, 2007), so the marginal contribution of m varies with $\mathcal{V}$ only through components of order at least two. We call the marginal contribution of m *context-dependent* if $(\Delta_m H_{\boldsymbol{x}})(\mathcal{V}) \neq (\Delta_m H_{\boldsymbol{x}})(\mathcal{W})$ for some $\mathcal{V}, \mathcal{W} \subseteq [M] \setminus \{m\}$.

Proposition 1 (Context dependence and higher-order interactions)**.** *For a fixed $H_{\boldsymbol{x}}$, some modality $m \in [M]$ has a context-dependent marginal contribution if and only if $H_{\boldsymbol{x}}$ has a higher-order interaction, that is, $\boldsymbol{p}_{\mathcal{S}} \neq \mathbf{0}$ for some $\mathcal{S} \subseteq [M]$ with $|\mathcal{S}| \geqslant 2$.*

A formal proof is given in Appendix A.1. This equivalence suggests that modelling how each modality's marginal contribution changes with its context can induce higher-order interactions without direct parameterisation, motivating the construction introduced in the next section.

3 METHOD

LATFUSE instantiates the set function in §2 through context-conditioned single-modality extensions (Figure 1). Modality-specific towers supply reusable representations (§3.1), and a shared refinement operator produces edge messages (§3.2). We average these messages to construct subset states (§3.3) and apply Möbius inversion before the learned readout (§3.4). The training objective supervises task and subset predictions, calibrates edge gates, and penalises component similarity (§3.5).

3.1 INTRA-MODALITY TOKEN MIXING

LATFUSE receives modality-specific encoder outputs $\boldsymbol{e}_m \in \mathbb{R}^{d_m^{\text{enc}} \times N}$ with N position-aligned tokens; task-specific resampling precedes fusion. For $\boldsymbol{S} \in \mathbb{R}^{d \times N}$, we write $\boldsymbol{S}[n] \in \mathbb{R}^d$ for its n-th column. A modality-specific token-mixing tower produces $\boldsymbol{x}_m \in \mathbb{R}^{d_x \times N}$:

$$\boldsymbol{x}_m := \text{Tower}_m\big(\boldsymbol{W}_m^{\text{enc}} \boldsymbol{e}_m + \boldsymbol{\pi}_x\big) + \boldsymbol{\mu}_m. \tag{3}$$

Here, $\boldsymbol{W}_m^{\text{enc}}$ projects to width d_x, and $\boldsymbol{\pi}_x$ encodes token positions. We add the learned modality embedding $\boldsymbol{\mu}_m$ to each output token to identify its source to the shared refinement operator. All covering edges that add m reuse $\boldsymbol{x}_m$. Appendix B specifies the tower architectures and configurations.

3.2 Edge Refinement Operator

Motivated by Proposition 1, we parameterise each single-modality extension as a context-conditioned refinement along a covering edge. For an edge $\mathcal{V} \xrightarrow{+m} \mathcal{U}$, the refinement operator f uses $\boldsymbol{h}_{\mathcal{V}}$ as context to incorporate $\boldsymbol{x}_m$. We initialise $\boldsymbol{z}^{(0)} \coloneqq \boldsymbol{h}_{\mathcal{V}}$ and apply f for T steps with $\boldsymbol{x}_m$ fixed, sharing parameters across edges and steps.

Context-conditioned cross-attention. At step $t = 0, \dots, T-1$, the edge state supplies queries and the added modality supplies keys and values:

$$\boldsymbol{q}^{(t)} \coloneqq \boldsymbol{W}^{\mathrm{Q}} \,\mathrm{LN}\Big(\boldsymbol{z}^{(t)} + \boldsymbol{\pi}_z\Big), \qquad \boldsymbol{k}_m \coloneqq \boldsymbol{W}^{\mathrm{K}} \boldsymbol{x}_m, \qquad \boldsymbol{v}_m \coloneqq \boldsymbol{W}^{\mathrm{V}} \boldsymbol{x}_m. \tag{4}$$

Here, $\boldsymbol{\pi}_z$ is a positional embedding, LN denotes layer normalisation, and all three projections output width d_k. The source neighbourhood $\mathcal{N}(n)$ is $\{n\}$ for pointwise attention, $\{j \in [N] : |j - n| \leqslant w\}$ for local attention with radius w, or $[N]$ for global attention.

We add a fixed null slot $\perp$ with $\boldsymbol{k}_m[\perp] \coloneqq \boldsymbol{0}, \boldsymbol{v}_m[\perp] \coloneqq \boldsymbol{0}$ and normalise attention over $\mathcal{N}^+(n) \coloneqq \mathcal{N}(n) \cup \{\perp\}$:

$$\boldsymbol{\alpha}^{(t)}[n, \cdot] \coloneqq \mathrm{softmax}\left(\frac{\boldsymbol{q}^{(t)}[n]^\top \boldsymbol{k}_m[\mathcal{N}^+(n)]}{\sqrt{d_k}}\right), \quad \boldsymbol{a}^{(t)}[n] \coloneqq \boldsymbol{W}^{\mathrm{O}} \sum_{j \in \mathcal{N}^+(n)} \alpha^{(t)}[n, j] \boldsymbol{v}_m[j]. \tag{5}$$

Here, $\boldsymbol{W}^{\mathrm{O}}$ projects to width d_h. The null slot allows the total content attention mass $1 - \alpha^{(t)}[n, \perp]$ to be smaller than one. Averaging this mass over positions and steps gives the edge gate in §3.5.

Identity-preserving refinement. We combine the attention output with a feedforward update:

$$\boldsymbol{z}^{(t+1)} \coloneqq f\Big(\boldsymbol{z}^{(t)}; \boldsymbol{x}_m\Big) = \boldsymbol{z}^{(t)} + \boldsymbol{a}^{(t)} + \phi\Big(\mathrm{LN}\Big(\boldsymbol{z}^{(t)} + \boldsymbol{a}^{(t)}\Big)\Big) - \phi\Big(\mathrm{LN}\Big(\boldsymbol{z}^{(t)}\Big)\Big). \tag{6}$$

Here, $\phi : \mathbb{R}^{d_h} \to \mathbb{R}^{d_h}$ is a shared two-layer positionwise feedforward network. When $\boldsymbol{a}^{(t)} = \boldsymbol{0}$, the feedforward terms cancel and $\boldsymbol{z}^{(t+1)}$ equals $\boldsymbol{z}^{(t)}$. After T steps, the edge message $\boldsymbol{z}_{\mathcal{V} \to \mathcal{U}} \coloneqq \boldsymbol{z}^{(T)}$ provides a candidate state for $\mathcal{U}$.

3.3 Subset State Construction over Covering Edges

Each incoming covering edge provides a context-specific candidate representation for its destination subset. We combine these candidates to define a unique subset state and thereby instantiate the set function $H_{\boldsymbol{x}}$. LatFuse constructs subset states in increasing order $r = 1, \dots, M$ by averaging edge messages from all direct predecessors:

$$\boldsymbol{h}_{\mathcal{U}} \coloneqq \frac{1}{|\mathcal{U}|} \sum_{m \in \mathcal{U}} \boldsymbol{z}_{\mathcal{U} \setminus \{m\} \to \mathcal{U}}, \qquad \mathcal{U} \in \mathcal{G}_r. \tag{7}$$

This average combines extensions for every choice of the last added modality. For a singleton $\mathcal{U} = \{m\}$, refinement starts from $\boldsymbol{h}_{\varnothing} = \boldsymbol{0}$. States of the same order can be computed in parallel, giving M sequential orders with T refinement steps per order. The aggregated states determine each marginal contribution $\Delta_e H_{\boldsymbol{x}} = \boldsymbol{h}_{\mathcal{U}} - \boldsymbol{h}_{\mathcal{V}}$.

Proposition 2 (Subset-restricted dependence)**.** *For fixed model parameters and any non-empty $\mathcal{U} \subseteq [M]$, $H_{\boldsymbol{x}}(\mathcal{U})$ depends only on $(\boldsymbol{x}_m)_{m \in \mathcal{U}}$.*

Induction over $|\mathcal{U}|$ gives the result (Appendix A.2). For observed modalities $\mathcal{O} \subseteq [M]$, the same recurrence constructs every state on $2^{\mathcal{O}}$ using only observed inputs.

3.4 Möbius Decomposition and Readout

Having constructed the set function $H_{\boldsymbol{x}}$, we return to the global interaction view of §2. We apply Möbius inversion to the constructed states to obtain interaction components $\boldsymbol{p}_{\mathcal{S}}$. By Proposition 2, each component depends only on $(\boldsymbol{x}_m)_{m \in \mathcal{S}}$.

Order aggregation. For $\mathcal{S} \in \mathcal{G}_r$, let $\boldsymbol{u}_\mathcal{S} \coloneqq \boldsymbol{W}_r^{\mathrm{mob}} \boldsymbol{p}_\mathcal{S}$, where $\boldsymbol{W}_r^{\mathrm{mob}} \in \mathbb{R}^{d_u \times d_h}$ is a bias-free projection shared across $\mathcal{G}_r$. Using a bias-free two-layer positionwise network $\chi : \mathbb{R}^{d_u} \to \mathbb{R}^{d_u}$ shared across subsets and orders, we aggregate

$$\boldsymbol{g}_r \coloneqq \frac{1}{\sqrt{|\mathcal{G}_r|}} \sum_{\mathcal{S} \in \mathcal{G}_r} (\boldsymbol{u}_\mathcal{S} + \chi(\boldsymbol{u}_\mathcal{S})) \in \mathbb{R}^{d_u \times N}. \tag{8}$$

We choose χ with $\chi(\mathbf{0}) = \mathbf{0}$, so zero components contribute zero to $\boldsymbol{g}_r$.

Cross-order residual fusion. Let $\psi : \mathbb{R}^{2d_u} \to \mathbb{R}^{d_u}$ be a two-layer positionwise network shared across orders, and let $[\cdot;\cdot]$ denote feature concatenation. Starting from $\boldsymbol{c}_0 \coloneqq \mathbf{0}$, we combine orders sequentially for $r = 1, \ldots, M$:

$$\boldsymbol{c}_r \coloneqq \boldsymbol{c}_{r-1} + \boldsymbol{g}_r + \psi(\mathrm{LN}([\boldsymbol{c}_{r-1}; \boldsymbol{g}_r])) - \psi(\mathrm{LN}([\boldsymbol{c}_{r-1}; \mathbf{0}])) \,. \tag{9}$$

When $\boldsymbol{g}_r = \mathbf{0}$, the two ψ terms cancel. An affine predictor $\boldsymbol{W}^{\mathrm{out}}(\cdot) + \boldsymbol{b}^{\mathrm{out}}$ maps the mean of $\boldsymbol{c}_M$ over positions to $\boldsymbol{o}^{\mathrm{task}}$ for global tasks, or each $\boldsymbol{c}_M[n]$ to $\boldsymbol{o}^{\mathrm{task}}[n]$ for positionwise tasks.

3.5 Training Objective

Let Y denote the target, $\ell_{\mathrm{task}}(\cdot, Y)$ the task loss, and $\mathbb{E}$ the expectation over training examples. We use cross-entropy for classification and task-specific regression losses, with details in Appendix B. We also encourage the edge gates and interaction components to retain their intended roles.

Task and subset supervision. A predictor $\mathrm{Pred}_{\mathrm{subset}}$ shared across subsets and orders assigns each non-empty $\mathcal{U} \subseteq [M]$ the loss $\ell_\mathcal{U} \coloneqq \ell_{\mathrm{task}}(\mathrm{Pred}_{\mathrm{subset}}(\boldsymbol{h}_\mathcal{U}), Y)$. We supervise the task output and subset states, giving each order equal weight:

$$\mathcal{L}_{\mathrm{main}} \coloneqq \mathbb{E}\big[\ell_{\mathrm{task}}\big(\boldsymbol{o}^{\mathrm{task}}, Y\big)\big] \,, \qquad \mathcal{L}_{\mathrm{subset}} \coloneqq \frac{1}{M} \sum_{r=1}^{M} \frac{1}{|\mathcal{G}_r|} \sum_{\mathcal{U} \in \mathcal{G}_r} \mathbb{E}[\ell_\mathcal{U}]. \tag{10}$$

Gate calibration. We use subset prediction losses to calibrate the attention assigned to the added modality in each context. Let $\ell_{\mathrm{prior}}(Y)$ denote the loss of a fixed, input-independent target prior predictor and $\varepsilon > 0$ a numerical stabiliser. With $a_\varnothing \coloneqq 0, A_\varnothing \coloneqq 0$, we define a clipped predictability proxy for each non-empty subset and its monotone envelope:

$$A_\mathcal{U} \coloneqq \max_{\mathcal{T} \subseteq \mathcal{U}} \mathrm{stopgrad}(a_\mathcal{T}), \quad \text{where} \quad a_\mathcal{U} \coloneqq \mathrm{clip}_{[0,1]}\left(1 - \frac{\ell_\mathcal{U}}{\ell_{\mathrm{prior}}(Y) + \varepsilon}\right). \tag{11}$$

The proxy estimates normalised improvement over the prior, while stopgrad blocks gradients through these estimates. For $e = (\mathcal{V}, \mathcal{U}) \in \mathcal{E}$, the envelope ensures $0 \leqslant \Delta_e A = A_\mathcal{U} - A_\mathcal{V} \leqslant A_\mathcal{U}$. We calibrate the mean content attention mass γ_e against this gain:

$$\mathcal{L}_{\mathrm{gate}} \coloneqq \mathbb{E}\left[\frac{\sum_{e \in \mathcal{E}} (A_\mathcal{U} \gamma_e - \Delta_e A)^2}{\sum_{e \in \mathcal{E}} A_\mathcal{U}^2 + \varepsilon}\right], \quad \text{where} \quad \gamma_e \coloneqq 1 - \frac{1}{TN} \sum_{t=0}^{T-1} \sum_{n=1}^{N} \alpha_e^{(t)}[n, \perp]. \tag{12}$$

Here, $\alpha_e^{(t)}$ denotes the attention weights on edge e. For $A_\mathcal{U} > 0$, the loss encourages γ_e to match the relative predictability gain $\Delta_e A / A_\mathcal{U}$. We use the predictability estimates only during training.

Component decorrelation. Let $K \coloneqq 2^M - 1$ be the number of components. We penalise the squared stabilised cosine similarity over all ordered pairs of distinct components:

$$\mathcal{L}_{\mathrm{decorr}} \coloneqq \mathbb{E}\left[\frac{1}{K(K-1)} \sum_{\mathcal{S}, \mathcal{S}' \in 2^{[M]} \setminus \{\varnothing\}, \mathcal{S} \neq \mathcal{S}'} \langle \widehat{\boldsymbol{p}}_\mathcal{S}, \widehat{\boldsymbol{p}}_{\mathcal{S}'} \rangle_{\mathrm{F}}^2\right], \quad \text{where} \quad \widehat{\boldsymbol{p}}_\mathcal{S} \coloneqq \frac{\boldsymbol{p}_\mathcal{S}}{\sqrt{\|\boldsymbol{p}_\mathcal{S}\|_{\mathrm{F}}^2 + \varepsilon}}. \tag{13}$$

Here, $\|\cdot\|_{\mathrm{F}}$ and $\langle \cdot, \cdot \rangle_{\mathrm{F}}$ denote the Frobenius norm and inner product, respectively.

Total objective. We minimise

$$\mathcal{L} \coloneqq \mathcal{L}_{\mathrm{main}} + \lambda_{\mathrm{subset}} \mathcal{L}_{\mathrm{subset}} + \lambda_{\mathrm{gate}} \mathcal{L}_{\mathrm{gate}} + \lambda_{\mathrm{decorr}} \mathcal{L}_{\mathrm{decorr}}. \tag{14}$$

The coefficients $\lambda_{\mathrm{subset}}$, λ_{gate}, and $\lambda_{\mathrm{decorr}}$ are fixed non-negative hyperparameters.

4 Related Work

We focus on prior work most directly related to our formulation. We further discuss broader connections to general fusion architectures, alternative interaction decompositions, missing-modality learning, and optimisation in Appendix C.

Explicit multimodal interactions. Tensor Fusion Network and Low-Rank Multimodal Fusion represent cross-modal interactions through tensor products or their factorised approximations (Zadeh et al., 2017; Liu et al., 2018). Dynamic Fusion Graph and Factorized Multimodal Transformer explicitly model connections or attention factors associated with modality combinations (Zadeh et al., 2018a; 2019), and Multimodal Residual Optimization organises subset predictors by interaction order (Wörtwein et al., 2022). These approaches parameterise modality products, modality combinations, or order-specific predictors directly. LatFuse instead parameterises shared context-conditioned single-modality extensions on the Boolean lattice, from which subset representations and their higher-order interaction components follow through composition and Möbius inversion.

Context-conditioned fusion. Gated Multimodal Units and Quality-aware Multimodal Fusion adapt modality contributions through input-dependent gates or uncertainty-aware weights (Ovalle et al., 2017; Zhang et al., 2023c), and MulT models directional cross-modal dependencies through attention (Tsai et al., 2019). Interaction-aware mixture-of-experts methods further combine specialised experts with sample-dependent routing (Yu et al., 2024; Xin et al., 2025). In LatFuse, the conditioning context has an explicit combinatorial meaning: each refinement edge adds one modality to the complete subset of modalities already present. This subset-level conditioning links local adaptation directly to the marginal contributions whose context dependence determines higher-order interactions, while the null attention slot and gate calibration control how much content each edge admits.

Interaction analysis and set functions. Empirical Multimodally-Additive Projection measures non-additive predictive interactions, while InterSHAP quantifies cross-modal interactions using Shapley interaction values (Hessel & Lee, 2020; Wenderoth et al., 2025). Partial Information Decomposition characterises redundancy, uniqueness, and synergy and has also motivated multimodal representation-learning objectives (Williams & Beer, 2010; Liang et al., 2023; Yang et al., 2026). LatFuse instead defines a representation-valued set function over modality subsets and relates its higher-order interaction components to context-dependent marginal contributions. This representation-level decomposition characterises deviations from additivity in learned subset states, whereas information-theoretic decompositions quantify how modalities share task-relevant information.

5 Experiments

We evaluate LatFuse on four representative multimodal benchmarks with three or four modalities, specifically TCGA-BRCA (Cancer Genome Atlas Research Network et al., 2013), MLOmics-Pan-cancer (Yang et al., 2025b), CMU-MOSI (Zadeh et al., 2016), and DeLiVER (Zhang et al., 2023b), spanning multiomics classification, sentiment analysis, and semantic segmentation (Table 1). Together, these benchmarks cover different configurations and prediction tasks, allowing us to assess how well the proposed fusion framework generalises across multimodal settings. Our experiments examine predictive performance, robustness to missing modalities, and recovery of known interaction structure.

Table 1: Benchmarks, prediction tasks, modalities, and evaluation metrics.

Dataset	Task	Modalities	Metrics
TCGA-BRCA	5-class classification	{mRNA, miRNA, DNA}	ACC, weighted F_1, macro F_1
MLOmics-Pan-cancer	32-class classification	{mRNA, miRNA, DNA, CNV}	ACC, w.F_1, m.F_1, PREC, NMI, ARI
CMU-MOSI	$[-3, 3]$ reg. and 7-class cls.	{Language, Acoustic, Visual}	MAE, Corr, ACC_7, ACC_2, F_1
DeLiVER	25-class segmentation	{RGB, Depth, Event, LiDAR}	mIoU

5.1 Datasets and Experimental Setup

TCGA-BRCA & MLOmics-Pan-cancer. For 5-class PAM50 subtype classification, we use the preprocessed features distributed with MOGONET (Wang et al., 2021) and repeat 5-fold cross-validation with 20 random seeds. For 32-class cancer type classification, we use the features from the MLOmics toolkit (Yang et al., 2025b) and repeat 5-fold cross-validation with 5 random seeds.

Table 2: Main results on TCGA-BRCA under the full-modality setting.

Method	ACC (%) ↑	weighted F_1 (%) ↑	macro F_1 (%) ↑	#Params (M) ↓	#FLOPs (G) ↓
MM-Dynamics (Han et al., 2022)	86.94±0.78	87.33±0.83	83.98±1.16	4.794	2.520
MOGLAM (Ouyang et al., 2023)	85.69±0.42	86.25±0.58	82.10±0.41	5.372	4.579
DPNET (Zou et al., 2023)	87.81±1.01	88.41±1.20	85.21±1.24	3.620	6.871
DeepMKL (Briscik et al., 2024)	85.63±1.83	85.69±2.17	82.04±2.32	1.646	**0.864**
DMIB (Fang et al., 2024)	86.22±0.49	86.44±0.51	82.30±0.98	6.535	2.248
MOSGAT (Wu et al., 2024)	87.17±0.41	87.11±0.54	84.08±0.63	2.522	7.822
CLCLSA (Zhao et al., 2024)	86.09±1.45	86.03±1.62	80.42±2.71	2.945	1.453
MOHGCN (Wu et al., 2025)	87.91±0.38	88.25±0.39	85.39±0.84	8.037	18.510
LATFUSE	**89.39±0.33**	**89.75±0.42**	**87.08±0.51**	**1.252**	4.039

Table 3: Main results on MLOmics-Pan-cancer under the full-modality setting.

Method	ACC (%) ↑	weighted F_1 (%) ↑	macro F_1 (%) ↑	PREC (%) ↑	NMI (%) ↑	ARI (%) ↑
MOGONET (Wang et al., 2021)	94.96±0.02	94.91±0.02	93.43±0.57	94.97±0.19	94.13±0.27	93.26±0.40
MM-Dynamics (Han et al., 2022)	95.64±0.07	95.58±0.08	92.40±0.21	95.66±0.08	95.05±0.04	94.12±0.09
CustOmics (Benkirane et al., 2023)	96.92±0.06	96.28±0.09	92.89±0.36	96.55±0.06	95.43±0.08	95.70±0.16
DMIB (Fang et al., 2024)	96.89±0.19	96.80±0.19	94.57±0.67	96.92±0.19	96.09±0.23	95.20±0.29
MOSGAT (Wu et al., 2024)	96.67±0.31	96.53±0.37	92.00±1.11	96.75±0.34	96.12±0.28	95.48±0.35
LATFUSE	**97.71±0.08**	**97.56±0.09**	**95.26±0.17**	**97.60±0.09**	**97.37±0.08**	**96.83±0.13**

Table 4: Main results on CMU-MOSI under the full-modality setting.

Method	MAE ↓	Corr ↑	ACC_7 (%) ↑	ACC_2 (*has0* / *non0*) (%) ↑	F_1 (*has0* / *non0*) (%) ↑
SUGRM (Hwang & Kim, 2023)	0.714±0.020	0.794±0.004	47.37±0.84	84.02±0.61 / 85.87±1.17	83.85±0.59 / 85.79±1.06
ALMT (Zhang et al., 2023a)	0.684±0.005	0.806±0.002	48.07±0.85	84.07±0.70 / 85.66±0.68	84.03±0.69 / 85.67±0.67
HSCL (Chen et al., 2024)	0.701±0.020	0.799±0.013	45.66±1.25	84.51±0.41 / 86.21±0.62	84.45±0.36 / 86.21±0.55
MSA-TeCaFN (Huang et al., 2024)	0.714±0.024	0.792±0.016	45.27±0.97	83.84±1.25 / 86.75±1.47	84.74±1.22 / 86.73±1.41
CLGSI (Yang et al., 2024)	0.714±0.039	0.797±0.014	45.99±2.11	83.43±0.83 / 85.90±1.15	84.40±0.80 / 85.93±1.08
AGS-SMoE (Chen et al., 2025a)	0.694±0.028	0.798±0.013	45.85±0.96	85.39±1.23 / 87.33±1.09	85.28±1.21 / 87.29±1.05
MMA (Chen et al., 2025b)	0.692±0.009	0.811±0.004	47.31±1.17	84.81±0.93 / 86.72±0.86	84.77±0.92 / 86.75±0.85
MFMB-Net (Tao et al., 2025)	0.733±0.025	0.780±0.004	44.10±1.30	83.08±1.15 / 84.59±1.18	83.05±1.10 / 84.62±1.13
DLF (Wang et al., 2025)	0.747±0.016	0.774±0.013	44.22±1.22	83.02±0.48 / 84.68±0.40	83.01±0.51 / 84.73±0.42
FeaDA (Yin et al., 2025)	0.724±0.020	0.792±0.016	44.31±1.35	84.19±0.94 / 86.09±1.11	84.13±0.91 / 86.09±1.09
MFON (Zhang et al., 2025)	0.748±0.020	0.771±0.005	43.12±1.15	83.46±0.95 / 84.90±1.03	83.43±0.97 / 84.93±1.00
DPDF-LQ (Zhou et al., 2025)	0.684±0.009	0.807±0.002	46.69±1.05	84.12±0.25 / 86.49±0.18	83.90±0.21 / 86.35±0.13
LATFUSE	**0.662±0.007**	**0.823±0.005**	**50.95±1.30**	**86.17±0.70 / 88.66±0.58**	**86.15±0.65 / 88.61±0.57**

CMU-MOSI. Language features come from a fine-tuned BERT-base model (Devlin et al., 2019), and we use the precomputed acoustic and visual features from the CMU Multimodal SDK (Zadeh et al., 2018b), aligned at the word level. The model predicts continuous sentiment scores in $[-3, 3]$, which we evaluate with regression and classification metrics over 5 random seeds. For binary evaluation, *has0* retains zero-score samples whereas *non0* excludes them.

DeLiVER. Following StitchFusion's data processing, feature extraction, and evaluation protocol (Li et al., 2025), we replace its attention-side modules in MultiAdapter with LATFUSE. We retain the remaining architecture and training protocol and evaluate over 3 random seeds.

Tables report the mean ± standard deviation over the runs specified above. In each column, the best result is in **bold** and the second best is underlined for the baseline comparisons. Complete dataset processing, training, and architecture details are provided in Appendix D.

5.2 MAIN RESULTS

We evaluate LATFUSE under the full-modality setting following the protocols in §5.1.

Multiomics classification. On TCGA-BRCA, Table 2 shows that LATFUSE outperforms all baselines with 89.39% accuracy, 89.75% weighted F_1, and 87.08% macro F_1 and the fewest parameters at a mid-range compute cost. On MLOmics-Pan-cancer, LATFUSE exceeds the best baseline on each of the six metrics by 0.68 to 1.25 points in the four-modality setting, as shown in Table 3.

Multimodal sentiment analysis. Table 4 reports the results on CMU-MOSI. For continuous sentiment prediction, LATFUSE achieves the lowest MAE of 0.662 and the highest Corr of 0.823 among the compared methods. LATFUSE gains a 2.88-point improvement on ACC_7, and under binary evaluations, it also outperforms the best baseline by 0.78 to 1.33 points.

Multimodal semantic segmentation. Table 5 reports full-modality performance on DeLiVER. The integration of LATFUSE into StitchFusion raises its mIoU on DeLiVER from 68.21 to 69.72, the best result overall, extending the gains to positionwise prediction.

Table 5: Main results on DeLiVER.

Method	Modalities	mIoU (%) ↑
CMNeXt (Zhang et al., 2023b)	RGB, D, E, L	66.33
StitchFusion (Li et al., 2025)	RGB, D, E, L	68.21
Mul-VMamba (Ni et al., 2026)	RGB, D, E, L	<u>68.99</u>
StitchFusion + LATFUSE	RGB, D, E, L	**69.72±0.30**

Across the four benchmarks, LATFUSE achieves the strongest mean performance on the reported task metrics under full-modality evaluation, spanning three- and four-modality settings as well as global and positionwise prediction. The consistency of these gains across task types and evaluation measures supports the predictive utility of the proposed fusion construction beyond a single application setting.

5.3 ABLATION STUDIES

To identify which components drive the empirical gains of LATFUSE, we ablate the model along three aspects of its design: subset construction, interaction readout, and training regularisation. **A1** removes context conditioning, while **A4** removes intermediate subset states, directly testing the two components that define how representations are constructed over the Boolean lattice. **A2** removes Möbius inversion from the task readout, and **A3** removes the identity-preserving corrections, isolating the choices that transform the learned subset structure into predictions. Finally, **L1** and **L2** remove gate calibration and component decorrelation, respectively, to assess the additional contribution of the training objectives beyond the architectural construction. We evaluate all variants on TCGA-BRCA, MLOmics-Pan-cancer, and CMU-MOSI using the same data splits and training protocol as the full model, with implementation details provided in Appendix D.

Task-level results. Among the tested changes, removing intermediate subset states and context conditioning causes the largest performance degradation, supporting their role as the central components of the fusion (Table 6). Appendix E reports the remaining task metrics and structural diagnostics.

Table 6: Ablation results under the full-modality setting.

Variant	TCGA-BRCA m.F_1 (%) ↑	Pan-cancer m.F_1 (%) ↑	CMU-MOSI ACC_7 (%) ↑
Full	**87.08**	**95.26**	**50.95**
(A1) w/o context conditioning	83.60 (−3.48)	92.59 (−2.67)	45.42 (−5.53)
(A2) w/o Möbius inversion	85.65 (−1.43)	93.37 (−1.89)	48.61 (−2.34)
(A3) w/o identity preservation	84.64 (−2.44)	93.14 (−2.12)	47.03 (−3.92)
(A4) w/o intermediate subsets	81.75 (−5.33)	91.18 (−4.08)	44.30 (−6.65)
(L1) w/o $\mathcal{L}_{\text{gate}}$	84.17 (−2.91)	92.79 (−2.47)	47.93 (−3.02)
(L2) w/o $\mathcal{L}_{\text{decorr}}$	85.83 (−1.25)	93.23 (−2.03)	49.11 (−1.84)

5.4 ROBUST INFERENCE UNDER MISSING MODALITIES

Subset-restricted dependence (Proposition 2) allows LATFUSE to restrict inference to the sublattice induced by the modalities available at test time. We therefore examine whether a model trained only on complete observations can retain predictive performance when one modality is missing at inference, without retraining or parameter updates. Dataset-level missingness removes the same modality from every sample and evaluates a fixed observed subset. Heterogeneous missingness varies the unavailable modality across samples and evaluates multiple observed subsets with the same model; we assign missing modalities with approximately equal frequencies and independently of target labels. All methods use complete observations during training and retain their trained parameters during evaluation. Appendix D specifies the missingness protocol and fusion-level inference procedures.

Table 7 compares LATFUSE and its StitchFusion integration against two representative baselines per benchmark, with changes in points from each method's full-modality score in parentheses. Across the 14 missing-modality conditions, our models achieve the highest score in 13 and the smallest drop from their respective full-modality scores in 10. The results show that the subset representations learned from complete training samples remain useful across different observed modality combinations without parameter updates. Complete metrics are provided in Appendix F.

5.5 ANALYSIS ON SYNTHETIC TASKS

We next examine whether the interaction structure exposed by the readout reflects the structure actually required by the target function. Because such ground-truth interaction structure is unavailable in

Table 7: Inference under dataset-level and heterogeneous missingness.

TCGA-BRCA (m.F_1, %↑)	Full	−mRNA	−miRNA	−DNA		Heterogeneous
MM-Dynamics (Han et al., 2022)	83.98	51.32 (−32.66)	76.84 (−7.14)	79.20 (−4.78)		71.05 (−12.93)
MOHGCN (Wu et al., 2025)	85.39	73.07 (−12.32)	82.57 (−2.82)	81.74 (−3.65)		79.20 (−6.19)
LATFUSE	**87.08**	**76.38 (−10.70)**	**85.42 (−1.66)**	**83.92 (−3.16)**		**82.77 (−4.31)**
Pan-cancer (m.F_1, %↑)	Full	−mRNA	−miRNA	−DNA	−CNV	Heterogeneous
DMIB (Fang et al., 2024)	94.57	91.70 (−2.87)	93.58 (**−0.99**)	91.84 (−2.73)	93.30 (−1.27)	92.85 (−1.72)
MOSGAT (Wu et al., 2024)	92.00	87.54 (−4.46)	87.86 (−4.14)	89.28 (−2.72)	91.05 (**−0.95**)	88.83 (−3.17)
LATFUSE	**95.26**	**92.49 (−2.77)**	**93.74** (−1.52)	**92.59 (−2.67)**	**93.79** (−1.47)	**93.58 (−1.68)**
DeLiVER (mIoU, %↑)	Full	−RGB	−Depth	−Event	−LiDAR	Heterogeneous
StitchFusion (Li et al., 2025)	68.21	52.52 (−15.69)	37.08 (−31.13)	66.89 (−1.32)	66.80 (−1.41)	55.11 (−13.10)
Mul-VMamba (Ni et al., 2026)	68.99	57.68 (−11.31)	33.96 (−35.03)	**68.98 (−0.01)**	68.62 (**−0.37**)	55.71 (−13.28)
StitchFusion + LATFUSE	**69.72**	**60.18 (−9.54)**	**41.88 (−27.84)**	68.64 (−1.08)	**68.82** (−0.90)	**59.63 (−10.09)**

real-world benchmarks, we construct synthetic tasks with known functional decompositions. For independent scalar inputs $s_m \sim \mathrm{Uniform}[-\sqrt{3}, \sqrt{3}]$, we compare a non-linear additive target with a mixed-order target containing first-order effects, the pairwise interaction $s_1 s_2$, and the three-way interaction $s_1 s_2 s_3$. We extract each readout-stage increment as $d_r(\boldsymbol{s}) := \boldsymbol{W}^{\mathrm{out}}(\boldsymbol{c}_r(\boldsymbol{s}) - \boldsymbol{c}_{r-1}(\boldsymbol{s}))$, which gives $\widehat{f}(\boldsymbol{s}) = \boldsymbol{b}^{\mathrm{out}} + \sum_{r=1}^{3} d_r(\boldsymbol{s})$. We then apply fANOVA (Sobol', 2001) to each d_r under the prescribed input distribution and use Sobol indices to measure the share of within-stage variance associated with each modality subset.

In the additive setting, singleton effects dominate all readout stages, while the absolute variance decreases sharply after d_1, from 0.716 at d_1 to 0.018 at d_2, indicating only small later-stage corrections. For the mixed target, the dominant functional order progresses from first order in d_1 to second order in d_2 and third order in d_3 (Figure 2). The prescribed pair $\{1, 2\}$ accounts for 62.59% of the variance in d_2, whereas the other pairs each remain below 1%. The three-way term accounts for 88.22% of the variance in d_3, alongside smaller lower-order residuals.

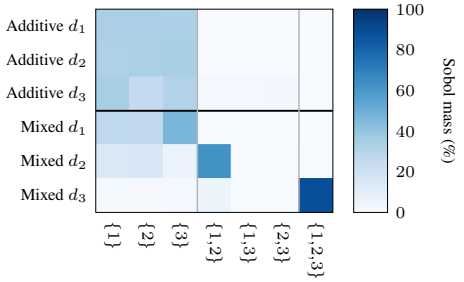


Figure 2: Subset-level Sobol mass.

These results show that the readout organises predictive effects according to the interaction structure of the target at both the order and subset levels. The remaining lower-order mass in later stages indicates that the stages are not a strict functional decomposition by order, but the dominant effects align with the prescribed interactions. Complete task and stagewise errors, numerical decomposition details, and conditional-response diagnostics (Goldstein et al., 2013) are provided in Appendix G.

6 Conclusion and Limitations

Conclusion. We introduced LATFUSE, a multimodal fusion module that constructs subset states on the Boolean lattice through shared edge refinement along single-modality extensions. Experiments on four benchmarks spanning multiomics classification, sentiment analysis, and semantic segmentation show consistent improvements over the compared baselines. Together, these results support subset states as a basis for effective multimodal fusion, with the same states enabling inference on observed subsets and analysis across interaction orders.

Limitations and future work. Our evaluation is limited to tasks with at most four modalities, and the complete Boolean lattice contains $2^M - 1$ non-empty subset states and therefore becomes increasingly expensive as M grows. Scaling to settings with many modalities may require sparse or low-order constructions that evaluate only a structured subset of interactions while retaining the benefits of subset-local computation. Our formulation also uses a Boolean lattice in which each modality has a binary inclusion state. Extending the framework to finite distributive lattices could support more structured notions of modality availability, but would require corresponding definitions of covering-edge refinement and interaction order.

AI USE STATEMENT

In this work, we *used* generative AI tools (including only large language models) to implement the code of our model and the training code for various experiments in §5.

We *have not used* generative AI tools to develop theoretical models or conceptual frameworks, to formulate mathematical claims, to provide critical ingredients for proving mathematical claims, to generate synthetic data sets, or to assist in the writing of proofs. Other subtasks requiring disclosure in the AI Policy for Authors are not applicable to this work.

Additionally, we *used* generative AI tools to implement the TikZ and Matplotlib code behind the figures. We *used* generative AI tools for sourcing and searching for information, identifying relevant literature, and summarising or analysing existing literature while surveying related work. We *used* generative AI tools to suggest a structure for the paper and edit the paper to improve readability.

We have reviewed all AI-assisted work. All AI-generated code was audited by the authors. Every source surfaced through AI-assisted literature search was retrieved and read in its original form. All citations were extracted manually from reliable sources.

We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have provided the experimental details in Appendix D and released the source code on https://anonymous.4open.science/r/LatFuse/.

REFERENCES

Padegal Amit, Omkar Mahesh Kashyap, Namitha Rayasam, Nidhi Shekhar, and Surabhi Narayan. Quantifying modality contributions via disentangling multimodal representations. *CoRR*, abs/2511.19470, 2025. doi: 10.48550/ARXIV.2511.19470. URL https://doi.org/10.48550/arXiv.2511.19470.

Hakim Benkirane, Yoann Pradat, Stefan Michiels, and Paul-Henry Cournède. Customics: A versatile deep-learning based strategy for multi-omics integration. *PLoS Comput. Biol.*, 19(3), 2023. doi: 10.1371/JOURNAL.PCBI.1010921. URL https://doi.org/10.1371/journal.pcbi.1010921.

Mitja Briscik, Gabriele Tazza, László Vidács, Marie-Agnès Dillies, and Sébastien Déjean. Supervised multiple kernel learning approaches for multi-omics data integration. *BioData Min.*, 17(1), 2024. doi: 10.1186/S13040-024-00406-9. URL https://doi.org/10.1186/s13040-024-00406-9.

Cancer Genome Atlas Research Network, John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas Pan-Cancer analysis project. *Nat. Genet.*, 45(10): 1113–1120, October 2013.

Jili Chen, Qionghao Huang, Changqin Huang, and Xiaodi Huang. Actual cause-guided adaptive gradient scaling for balanced multimodal sentiment analysis. *ACM Trans. Multim. Comput. Commun. Appl.*, 21(6):179:1–179:24, 2025a. doi: 10.1145/3736415. URL https://doi.org/10.1145/3736415.

Kezhou Chen, Shuo Wang, and Yanbin Hao. Hierarchical supervised contrastive learning for multimodal sentiment analysis. In Stevan Rudinac, Alan Hanjalic, Cynthia C. S. Liem, Marcel Worring, Björn Þór Jónsson, Bei Liu, and Yoko Yamakata (eds.), *MultiMedia Modeling - 30th International Conference, MMM 2024, Amsterdam, The Netherlands, January 29 - February 2, 2024, Proceedings, Part II*, volume 14555 of *Lecture Notes in Computer Science*, pp. 56–69. Springer, 2024. doi: 10.1007/978-3-031-53308-2_5. URL https://doi.org/10.1007/978-3-031-53308-2_5.

Kezhou Chen, Shuo Wang, Huixia Ben, Shengeng Tang, and Yanbin Hao. Mixture of multimodal adapters for sentiment analysis. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pp. 1822–1833. Association for Computational Linguistics, 2025b. doi: 10.18653/V1/2025.NAACL-LONG.90. URL https://doi.org/10.18653/v1/2025.naacl-long.90.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1423. URL https://doi.org/10.18653/v1/n19-1423.

Yingying Fang, Shuang Wu, Sheng Zhang, Chaoyan Huang, Tieyong Zeng, Xiaodan Xing, Simon Walsh, and Guang Yang. Dynamic multimodal information bottleneck for multimodality classification. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024*, pp. 7681–7691. IEEE, 2024. doi: 10.1109/WACV57701.2024.00752. URL https://doi.org/10.1109/WACV57701.2024.00752.

Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24, 09 2013. doi: 10.1080/10618600.2014.907095.

Michel Grabisch and Christophe Labreuche. Derivative of functions over lattices as a basis for the notion of interaction between attributes. *Ann. Math. Artif. Intell.*, 49(1-4):151–170, 2007. doi: 10.1007/S10472-007-9052-7. URL https://doi.org/10.1007/s10472-007-9052-7.

Zongbo Han, Fan Yang, Junzhou Huang, Changqing Zhang, and Jianhua Yao. Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 20675–20685. IEEE, 2022. doi: 10.1109/CVPR52688.2022.02005. URL https://doi.org/10.1109/CVPR52688.2022.02005.

Mohammad Havaei, Nicolas Guizard, Nicolas Chapados, and Yoshua Bengio. Hemis: Hetero-modal image segmentation. In Sébastien Ourselin, Leo Joskowicz, Mert R. Sabuncu, Gözde B. Ünal, and William M. Wells III (eds.), *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016 - 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II*, volume 9901 of *Lecture Notes in Computer Science*, pp. 469–477, 2016. doi: 10.1007/978-3-319-46723-8_54. URL https://doi.org/10.1007/978-3-319-46723-8_54.

Jack Hessel and Lillian Lee. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 861–877. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.EMNLP-MAIN.62. URL https://doi.org/10.18653/v1/2020.emnlp-main.62.

Qionghao Huang, Jili Chen, Changqin Huang, Xiaodi Huang, and Yi Wang. Text-centered cross-sample fusion network for multimodal sentiment analysis. *Multim. Syst.*, 30(4):228, 2024. doi: 10.1007/S00530-024-01421-W. URL https://doi.org/10.1007/s00530-024-01421-w.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning? (provably). In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9226–9259. PMLR, 2022. URL https://proceedings.mlr.press/v162/huang22e.html.

Yewon Hwang and Jong-Hwan Kim. Self-supervised unimodal label generation strategy using recalibrated modality representations for multimodal sentiment analysis. In Andreas Vlachos and Isabelle Augenstein (eds.), *Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, volume EACL 2023 of *Findings of ACL*, pp. 35–46. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-EACL.2. URL https://doi.org/10.18653/v1/2023.findings-eacl.2.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and João Carreira. Perceiver: General perception with iterative attention. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4651–4664. PMLR, 2021. URL http://proceedings.mlr.press/v139/jaegle21a.html.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey E. Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019. URL http://proceedings.mlr.press/v97/kornblith19a.html.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753. PMLR, 2019. URL http://proceedings.mlr.press/v97/lee19d.html.

Bingyu Li, Da Zhang, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. Stitchfusion: Weaving any visual modalities to enhance multimodal semantic segmentation. In Cathal Gurrin, Klaus Schoeffmann, Min Zhang, Luca Rossetto, Stevan Rudinac, Duc-Tien Dang-Nguyen, Wen-Huang Cheng, Phoebe Chen, and Jenny Benois-Pineau (eds.), *Proceedings of the 33rd ACM International Conference on Multimedia, MM 2025, Dublin, Ireland, October 27-31, 2025*, pp. 1308–1317. ACM, 2025. doi: 10.1145/3746027.3755110. URL https://doi.org/10.1145/3746027.3755110.

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard J. Chen, Zihao Deng, Nicholas B. Allen, Randy Auerbach, Faisal Mahmood, Russ Salakhutdinov, and Louis-Philippe Morency. Quantifying & modeling multimodal interactions: An information decomposition framework. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/575286a73f238b6516ce0467d67eadb2-Abstract-Conference.html.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pp. 2247–2256. Association for Computational Linguistics, 2018. doi: 10.18653/V1/P18-1209. URL https://aclanthology.org/P18-1209/.

Divyam Madaan, Varshan Muhunthan, Kyunghyun Cho, and Sumit Chopra. Multi-modal data spectrum: Multi-modal datasets are multi-dimensional. *CoRR*, abs/2509.23499, 2025. doi: 10.48550/ARXIV.2509.23499. URL https://doi.org/10.48550/arXiv.2509.23499.

Natalia Neverova, Christian Wolf, Graham W. Taylor, and Florian Nebout. Moddrop: Adaptive multi-modal gesture recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(8):1692–1706, 2016. doi: 10.1109/TPAMI.2015.2461544. URL https://doi.org/10.1109/TPAMI.2015.2461544.

Rongrong Ni, Yuanhui Guo, Biao Yang, Yi Liu, Hai Wang, and Chuan Hu. Mul-vmamba: Multimodal semantic segmentation using selection-fusion-based vision-mamba. *Knowl. Based Syst.*, 334: 115119, 2026. doi: 10.1016/J.KNOSYS.2025.115119. URL https://doi.org/10.1016/j.knosys.2025.115119.

Dong Ouyang, Yong Liang, Le Li, Ning Ai, Shanghui Lu, Mingkun Yu, Xiaoying Liu, and Shengli Xie. Integration of multi-omics data using adaptive graph learning and attention mechanism for patient classification and biomarker identification. *Comput. Biol. Medicine*, 164:107303, 2023. doi: 10.1016/J.COMPBIOMED.2023.107303. URL https://doi.org/10.1016/j.compbiomed.2023.107303.

John Edison Arevalo Ovalle, Thamar Solorio, Manuel Montes-y-Gómez, and Fabio A. González. Gated multimodal units for information fusion. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=S12_nquOe.

Letitia Parcalabescu and Anette Frank. MM-SHAP: A performance-agnostic metric for measuring multimodal contributions in vision and language models & tasks. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 4032–4059. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.223. URL https://doi.org/10.18653/v1/2023.acl-long.223.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 8228–8237. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00806. URL https://doi.org/10.1109/CVPR52688.2022.00806.

Wasifur Rahman, Md. Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Mohammed E. Hoque. Integrating multimodal information in large pretrained transformers. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 2359–2369. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.ACL-MAIN.214. URL https://doi.org/10.18653/v1/2020.acl-main.214.

Adrià Recasens, Jason Lin, João Carreira, Andrew Jaegle, Luyu Wang, Jean-Baptiste Alayrac, Pauline Luc, Antoine Miech, Lucas Smaira, Ross Hemsley, and Andrew Zisserman. Zorro: the masked multimodal transformer. *CoRR*, abs/2301.09595, 2023. doi: 10.48550/ARXIV.2301.09595. URL https://doi.org/10.48550/arXiv.2301.09595.

Gian Carlo Rota. On the foundations of combinatorial theory i. theory of möbius functions. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 2(4):340–368, 1964. doi: 10.1007/BF00531932. URL https://doi.org/10.1007/BF00531932.

Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogério Feris, David Harwath, James R. Glass, and Hilde Kuehne. Everything at once - multi-modal fusion transformer for video retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 19988–19997. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01939. URL https://doi.org/10.1109/CVPR52688.2022.01939.

I.M Sobol'. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and Computers in Simulation*, 55(1):271–280, 2001. ISSN 0378-4754. doi: https://doi.org/10.1016/S0378-4754(00)00270-6. URL https://www.sciencedirect.com/science/article/pii/S0378475400002706. The Second IMACS Seminar on Monte Carlo Methods.

Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The shapley taylor interaction index. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9259–9268. PMLR, 2020. URL http://proceedings.mlr.press/v119/sundararajan20a.html.

Thomas M. Sutter, Imant Daunhawer, and Julia E. Vogt. Generalized multimodal ELBO. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria,*

May 3-7, 2021. OpenReview.net, 2021. URL https://openreview.net/forum?id=5Y21V0RDBV.

Chuanqi Tao, Jiaming Li, Tianzi Zang, and Peng Gao. A multi-focus-driven multi-branch network for robust multimodal sentiment analysis. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 1547–1555. AAAI Press, 2025. doi: 10.1609/AAAI.V39I2.32146. URL https://doi.org/10.1609/aaai.v39i2.32146.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In Anna Korhonen, David R. Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 6558–6569. Association for Computational Linguistics, 2019. doi: 10.18653/V1/P19-1656. URL https://doi.org/10.18653/v1/p19-1656.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Pan Wang, Qiang Zhou, Yawen Wu, Tianlong Chen, and Jingtong Hu. DLF: disentangled-language-focused multimodal sentiment analysis. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 21180–21188. AAAI Press, 2025. doi: 10.1609/AAAI.V39I20.35416. URL https://doi.org/10.1609/aaai.v39i20.35416.

Tongxin Wang, Wei Shao, Zhi Huang, Haixu Tang, Jie Zhang, Zhengming Ding, and Kun Huang. Mogonet integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nature Communications*, 12(1):3445, 2021. doi: 10.1038/s41467-021-23774-w. URL https://doi.org/10.1038/s41467-021-23774-w.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 12692–12702. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.01271. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Wang_What_Makes_Training_Multi-Modal_Classification_Networks_Hard_CVPR_2020_paper.html.

Laura Wenderoth, Konstantin Hemker, Nikola Simidjievski, and Mateja Jamnik. Measuring cross-modal interactions in multimodal models. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 21501–21509. AAAI Press, 2025. doi: 10.1609/AAAI.V39I20.35452. URL https://doi.org/10.1609/aaai.v39i20.35452.

Paul L. Williams and Randall D. Beer. Nonnegative decomposition of multivariate information. *CoRR*, abs/1004.2515, 2010. URL http://arxiv.org/abs/1004.2515.

Torsten Wörtwein, Lisa Sheeber, Nicholas B. Allen, Jeffrey F. Cohn, and Louis-Philippe Morency. Beyond additive fusion: Learning non-additive multimodal interactions. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, volume EMNLP 2022 of *Findings of ACL*, pp. 4681–4696. Association for Computational Linguistics, 2022. doi: 10.

18653/V1/2022.FINDINGS-EMNLP.344. URL https://doi.org/10.18653/v1/2022.findings-emnlp.344.

Mike Wu and Noah D. Goodman. Multimodal generative models for scalable weakly-supervised learning. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 5580–5590, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/1102a326d5f7c9e04fc3c89d0ede88c9-Abstract.html.

Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J. Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 24043–24055. PMLR, 2022. URL https://proceedings.mlr.press/v162/wu22d.html.

Wenhao Wu, Shudong Wang, Yuanyuan Zhang, Wenjing Yin, Yawu Zhao, and Shanchen Pang. MOSGAT: uniting specificity-aware gats and cross modal-attention to integrate multi-omics data for disease diagnosis. *IEEE J. Biomed. Health Informatics*, 28(9):5624–5637, 2024. doi: 10.1109/JBHI.2024.3415641. URL https://doi.org/10.1109/JBHI.2024.3415641.

Wenhao Wu, Shudong Wang, Yuanyuan Zhang, Kuijie Zhang, Wenjing Yin, and Shanchen Pang. MOHGCN: A trustworthy multi-omics data integration framework based on specificity-aware heterogeneous graph convolutional neural networks for disease diagnosis. *Expert Syst. Appl.*, 263: 125772, 2025. doi: 10.1016/J.ESWA.2024.125772. URL https://doi.org/10.1016/j.eswa.2024.125772.

Jiayi Xin, Sukwon Yun, Jie Peng, Inyoung Choi, Jenna L. Ballard, Tianlong Chen, and Qi Long. I2moe: Interpretable multimodal interaction-aware mixture-of-experts. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/xin25c.html.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10524–10533. PMLR, 2020. URL http://proceedings.mlr.press/v119/xiong20b.html.

Yang Yang, Xunde Dong, and Yupeng Qiang. CLGSI: A multimodal sentiment analysis framework based on contrastive learning guided by sentiment intensity. In Kevin Duh, Helena Gómez-Adorno, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, volume NAACL 2024 of *Findings of ACL*, pp. 2099–2110. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-NAACL.135. URL https://doi.org/10.18653/v1/2024.findings-naacl.135.

Zequn Yang, Hongfa Wang, and Di Hu. Efficient quantification of multimodal interaction at sample level. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025a. URL https://proceedings.mlr.press/v267/yang25aj.html.

Zequn Yang, Yake Wei, Haotian Ni, Zhihao Xu, and Di Hu. Information-theoretic decomposition for multimodal interaction learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2026, Denver, CO, USA, June 3-7, 2026*, pp. 30278–30287. Computer Vision Foundation, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Yang_Information-Theoretic_Decomposition_for_Multimodal_Interaction_Learning_CVPR_2026_paper.html.

Ziwei Yang, Rikuto Kotoge, Xihao Piao, Zheng Chen, Lingwei Zhu, Peng Gao, Yasuko Matsubara, Yasushi Sakurai, and Jimeng Sun. Mlomics: Cancer multi-omics database for machine learning. *Scientific Data*, 12(1):913, 2025b. doi: 10.1038/s41597-025-05235-x. URL https://doi.org/10.1038/s41597-025-05235-x.

Dapeng Yin, Bingxuan Hou, Mengna Gao, Shuyue Zhu, and Jun-li Wang. Feature decomposition-augmentation network for multimodal sentiment analysis. In Kentaro Inui, Sakriani Sakti, Haofen Wang, Derek F. Wong, Pushpak Bhattacharyya, Biplab Banerjee, Asif Ekbal, Tanmoy Chakraborty, and Dhirendra Pratap Singh (eds.), *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, IJCNLP-AACL 2025, Mumbai, India, December 20-24, 2025*, pp. 86–98. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.IJCNLP-LONG.6. URL https://doi.org/10.18653/v1/2025.ijcnlp-long.6.

Haofei Yu, Zhengyang Qi, Lawrence Jang, Russ Salakhutdinov, Louis-Philippe Morency, and Paul Pu Liang. Mmoe: Enhancing multimodal models with mixtures of multimodal interaction experts. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 10006–10030. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.558. URL https://doi.org/10.18653/v1/2024.emnlp-main.558.

Sukwon Yun, Inyoung Choi, Jie Peng, Yangfan Wu, Jingxuan Bao, Qiyiwen Zhang, Jiayi Xin, Qi Long, and Tianlong Chen. Flex-moe: Modeling arbitrary modality combination via the flexible mixture-of-experts. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/b2f2af5403042b1344f4e93b35fb67d9-Abstract-Conference.html.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *CoRR*, abs/1606.06259, 2016. URL http://arxiv.org/abs/1606.06259.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pp. 1103–1114. Association for Computational Linguistics, 2017. doi: 10.18653/V1/D17-1115. URL https://doi.org/10.18653/v1/d17-1115.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pp. 2236–2246. Association for Computational Linguistics, 2018a. doi: 10.18653/V1/P18-1208. URL https://aclanthology.org/P18-1208/.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. Multi-attention recurrent network for human communication comprehension. In Sheila A. McIlraith and Kilian Q. Weinberger (eds.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 5642–5649. AAAI Press, 2018b. doi: 10.1609/AAAI.V32I1.12024. URL https://doi.org/10.1609/aaai.v32i1.12024.

Amir Zadeh, Chengfeng Mao, Kelly Shi, Yiwei Zhang, Paul Pu Liang, Soujanya Poria, and Louis-Philippe Morency. Factorized multimodal transformer for multimodal sequential learning. *CoRR*, abs/1911.09826, 2019. URL http://arxiv.org/abs/1911.09826.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 3391–3401, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/f22e4747da1aa27e363d86d40ff442fe-Abstract.html.

Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 756–767. Association for Computational Linguistics, 2023a. doi: 10.18653/V1/2023.EMNLP-MAIN.49. URL https://doi.org/10.18653/v1/2023.emnlp-main.49.

Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Delivering arbitrary-modal semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 1136–1147. IEEE, 2023b. doi: 10.1109/CVPR52729.2023.00116. URL https://doi.org/10.1109/CVPR52729.2023.00116.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 41753–41769. PMLR, 2023c. URL https://proceedings.mlr.press/v202/zhang23ar.html.

Xiangmin Zhang, Wei Wei, and Shihao Zou. Modal feature optimization network with prompt for multimodal sentiment analysis. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pp. 4611–4621. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.coling-main.309/.

Chen Zhao, Anqi Liu, Xiao Zhang, Xuewei Cao, Zhengming Ding, Qiuying Sha, Hui Shen, Hong-Wen Deng, and Weihua Zhou. CLCLSA: cross-omics linked embedding with contrastive learning and self attention for integration with incomplete multi-omics data. *Comput. Biol. Medicine*, 170: 108058, 2024. doi: 10.1016/J.COMPBIOMED.2024.108058. URL https://doi.org/10.1016/j.compbiomed.2024.108058.

Miao Zhou, Lina Yang, Thomas Wu, Dongnan Yang, and Xinru Zhang. Dual-path dynamic fusion with learnable query for multimodal sentiment analysis. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 11355–11365. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.571. URL https://doi.org/10.18653/v1/2025.emnlp-main.571.

Xin Zou, Chang Tang, Xiao Zheng, Zhenglai Li, Xiao He, Shan An, and Xinwang Liu. DPNET: dynamic poly-attention network for trustworthy multi-modal classification. In Abdulmotaleb El Saddik, Tao Mei, Rita Cucchiara, Marco Bertini, Diana Patricia Tobon Vallejo, Pradeep K. Atrey, and M. Shamim Hossain (eds.), *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pp. 3550–3559. ACM, 2023. doi: 10.1145/3581783.3612652. URL https://doi.org/10.1145/3581783.3612652.

A Propositions and Proofs

A.1 Context Dependence and Higher-Order Interactions

Preliminaries. Throughout this section, let $\mathcal{R}$ be a finite set, let $\mathcal{Y}$ be a real vector space, and let $\mathbf{0}$ denote the additive identity of $\mathcal{Y}$. For a set function $F : 2^{\mathcal{R}} \to \mathcal{Y}$, define its Möbius transform $\mu_F : 2^{\mathcal{R}} \to \mathcal{Y}$ by

$$\mu_F(\mathcal{S}) \coloneqq \sum_{\mathcal{T} \subseteq \mathcal{S}} (-1)^{|\mathcal{S} \setminus \mathcal{T}|} F(\mathcal{T}), \qquad \mathcal{S} \subseteq \mathcal{R}. \tag{15}$$

For the set function $H_{\boldsymbol{x}}$ in §2, the components defined in Equation (2) satisfy $\boldsymbol{p}_{\mathcal{S}} = \mu_{H_{\boldsymbol{x}}}(\mathcal{S})$ for every non-empty $\mathcal{S} \subseteq [M]$.

Lemma 1 (Möbius inversion)**.** *For any set function $F : 2^{\mathcal{R}} \to \mathcal{Y}$ and any $\mathcal{U} \subseteq \mathcal{R}$,*

$$F(\mathcal{U}) = \sum_{\mathcal{S} \subseteq \mathcal{U}} \mu_F(\mathcal{S}). \tag{16}$$

Proof. By the definition of μ_F and a change in the order of summation,

$$\sum_{\mathcal{S} \subseteq \mathcal{U}} \mu_F(\mathcal{S}) = \sum_{\mathcal{S} \subseteq \mathcal{U}} \sum_{\mathcal{T} \subseteq \mathcal{S}} (-1)^{|\mathcal{S} \setminus \mathcal{T}|} F(\mathcal{T}) = \sum_{\mathcal{T} \subseteq \mathcal{U}} F(\mathcal{T}) \sum_{\mathcal{T} \subseteq \mathcal{S} \subseteq \mathcal{U}} (-1)^{|\mathcal{S} \setminus \mathcal{T}|}. \tag{17}$$

For fixed $\mathcal{T} \subseteq \mathcal{U}$, every $\mathcal{S}$ with $\mathcal{T} \subseteq \mathcal{S} \subseteq \mathcal{U}$ has the unique form $\mathcal{S} = \mathcal{T} \cup \mathcal{Q}$ for some $\mathcal{Q} \subseteq \mathcal{U} \setminus \mathcal{T}$. Hence,

$$\sum_{\mathcal{T} \subseteq \mathcal{S} \subseteq \mathcal{U}} (-1)^{|\mathcal{S} \setminus \mathcal{T}|} = \sum_{\mathcal{Q} \subseteq \mathcal{U} \setminus \mathcal{T}} (-1)^{|\mathcal{Q}|}. \tag{18}$$

The binomial theorem gives a value of 1 when $\mathcal{T} = \mathcal{U}$ and 0 otherwise. Therefore, only the term $\mathcal{T} = \mathcal{U}$ remains, which gives $\sum_{\mathcal{S} \subseteq \mathcal{U}} \mu_F(\mathcal{S}) = F(\mathcal{U})$. □

Lemma 2 (Characterisation of constant set functions)**.** *A set function $F : 2^{\mathcal{R}} \to \mathcal{Y}$ is constant if and only if $\mu_F(\mathcal{S}) = \mathbf{0}$ for every non-empty $\mathcal{S} \subseteq \mathcal{R}$.*

Proof. Suppose first that $F(\mathcal{T}) = \boldsymbol{c}$ for every $\mathcal{T} \subseteq \mathcal{R}$ and some $\boldsymbol{c} \in \mathcal{Y}$. For every non-empty $\mathcal{S} \subseteq \mathcal{R}$,

$$\mu_F(\mathcal{S}) = \boldsymbol{c} \sum_{\mathcal{T} \subseteq \mathcal{S}} (-1)^{|\mathcal{S} \setminus \mathcal{T}|} = \mathbf{0}, \tag{19}$$

where the last equality follows from the binomial theorem. Conversely, suppose that $\mu_F(\mathcal{S}) = \mathbf{0}$ for every non-empty $\mathcal{S} \subseteq \mathcal{R}$. By Lemma 1, every $\mathcal{U} \subseteq \mathcal{R}$ satisfies

$$F(\mathcal{U}) = \sum_{\mathcal{S} \subseteq \mathcal{U}} \mu_F(\mathcal{S}) = \mu_F(\varnothing) = F(\varnothing). \tag{20}$$

Thus, F is constant. □

Lemma 3 (Möbius transform of a marginal contribution)**.** *Let $F : 2^{\mathcal{R}} \to \mathcal{Y}$ and fix $m \in \mathcal{R}$. Define the marginal set function $\Delta_m F : 2^{\mathcal{R} \setminus \{m\}} \to \mathcal{Y}$ by $(\Delta_m F)(\mathcal{V}) \coloneqq F(\mathcal{V} \cup \{m\}) - F(\mathcal{V})$. Then, for every $\mathcal{V} \subseteq \mathcal{R} \setminus \{m\}$,*

$$\mu_{\Delta_m F}(\mathcal{V}) = \mu_F(\mathcal{V} \cup \{m\}). \tag{21}$$

Proof. Fix $\mathcal{V} \subseteq \mathcal{R} \setminus \{m\}$. Every subset of $\mathcal{V} \cup \{m\}$ either excludes m and has the form $\mathcal{T}$, or includes m and has the form $\mathcal{T} \cup \{m\}$, for a unique $\mathcal{T} \subseteq \mathcal{V}$. The two corresponding terms in $\mu_F(\mathcal{V} \cup \{m\})$ have opposite signs because

$$|(\mathcal{V} \cup \{m\}) \setminus \mathcal{T}| = |\mathcal{V} \setminus \mathcal{T}| + 1 \quad \text{and} \quad |(\mathcal{V} \cup \{m\}) \setminus (\mathcal{T} \cup \{m\})| = |\mathcal{V} \setminus \mathcal{T}|. \tag{22}$$

Grouping these terms therefore gives

$$\begin{aligned} \mu_F(\mathcal{V} \cup \{m\}) &= \sum_{\mathcal{T} \subseteq \mathcal{V}} (-1)^{|\mathcal{V} \setminus \mathcal{T}|} \left(F(\mathcal{T} \cup \{m\}) - F(\mathcal{T})\right) \\ &= \sum_{\mathcal{T} \subseteq \mathcal{V}} (-1)^{|\mathcal{V} \setminus \mathcal{T}|} (\Delta_m F)(\mathcal{T}) = \mu_{\Delta_m F}(\mathcal{V}). \end{aligned} \tag{23}$$

□

Proposition 1 (Context dependence and higher-order interactions)**.** *For a fixed $H_{\boldsymbol{x}}$, some modality $m \in [M]$ has a context-dependent marginal contribution if and only if $H_{\boldsymbol{x}}$ has a higher-order interaction, that is, $\boldsymbol{p}_{\mathcal{S}} \neq \mathbf{0}$ for some $\mathcal{S} \subseteq [M]$ with $|\mathcal{S}| \geqslant 2$.*

Proof. Fix $m \in [M]$ and define $G_m \coloneqq \Delta_m H_{\boldsymbol{x}}$ as a set function on $2^{[M]\setminus\{m\}}$. By Lemma 2, G_m is constant if and only if $\mu_{G_m}(\mathcal{V}) = \mathbf{0}$ for every non-empty $\mathcal{V} \subseteq [M] \setminus \{m\}$. By Lemma 3,

$$\mu_{G_m}(\mathcal{V}) = \mu_{H_{\boldsymbol{x}}}(\mathcal{V} \cup \{m\}) = \boldsymbol{p}_{\mathcal{V}\cup\{m\}}, \tag{24}$$

where the last equality holds because $\mathcal{V} \cup \{m\}$ is non-empty. Consequently, the marginal contribution of modality m is context-independent if and only if

$$\boldsymbol{p}_{\mathcal{V}\cup\{m\}} = \mathbf{0}, \qquad \forall\, \varnothing \neq \mathcal{V} \subseteq [M] \setminus \{m\}. \tag{25}$$

The set function $H_{\boldsymbol{x}}$ has no context-dependent marginal contribution if and only if this condition holds for every $m \in [M]$. Every subset $\mathcal{S} \subseteq [M]$ with $|\mathcal{S}| \geqslant 2$ can be written as $\mathcal{S} = \mathcal{V} \cup \{m\}$ by choosing any $m \in \mathcal{S}$ and setting $\mathcal{V} = \mathcal{S} \setminus \{m\}$, which is non-empty. Conversely, every set $\mathcal{V} \cup \{m\}$ with $\varnothing \neq \mathcal{V} \subseteq [M] \setminus \{m\}$ has cardinality at least two. Hence, $H_{\boldsymbol{x}}$ has no context-dependent marginal contribution if and only if $\boldsymbol{p}_{\mathcal{S}} = \mathbf{0}$ for every $\mathcal{S} \subseteq [M]$ with $|\mathcal{S}| \geqslant 2$. Negating both equivalent statements proves the claim. □

Corollary (Reconstruction and marginal decomposition)**.** *For every $\mathcal{U} \subseteq [M]$, the subset state satisfies $\boldsymbol{h}_{\mathcal{U}} = \sum_{\varnothing \neq \mathcal{S} \subseteq \mathcal{U}} \boldsymbol{p}_{\mathcal{S}}$. For every covering edge $e : \mathcal{V} \xrightarrow{+m} \mathcal{V} \cup \{m\}$, its marginal contribution satisfies $\Delta_e H_{\boldsymbol{x}} = \sum_{\mathcal{T} \subseteq \mathcal{V}} \boldsymbol{p}_{\mathcal{T}\cup\{m\}}$.*

Proof. Applying Lemma 1 to $H_{\boldsymbol{x}}$ gives $\boldsymbol{h}_{\mathcal{U}} = \sum_{\mathcal{S} \subseteq \mathcal{U}} \mu_{H_{\boldsymbol{x}}}(\mathcal{S})$. The boundary condition $\boldsymbol{h}_{\varnothing} = \mathbf{0}$ gives $\mu_{H_{\boldsymbol{x}}}(\varnothing) = \mathbf{0}$, while $\mu_{H_{\boldsymbol{x}}}(\mathcal{S}) = \boldsymbol{p}_{\mathcal{S}}$ for every non-empty $\mathcal{S}$, which proves the reconstruction identity. For a covering edge $e : \mathcal{V} \xrightarrow{+m} \mathcal{V} \cup \{m\}$, Möbius inversion gives

$$\begin{aligned} \Delta_e H_{\boldsymbol{x}} = H_{\boldsymbol{x}}(\mathcal{V} \cup \{m\}) - H_{\boldsymbol{x}}(\mathcal{V}) &= \sum_{\mathcal{S} \subseteq \mathcal{V}\cup\{m\}} \mu_{H_{\boldsymbol{x}}}(\mathcal{S}) - \sum_{\mathcal{S} \subseteq \mathcal{V}} \mu_{H_{\boldsymbol{x}}}(\mathcal{S}) \\ &= \sum_{\mathcal{T} \subseteq \mathcal{V}} \mu_{H_{\boldsymbol{x}}}(\mathcal{T} \cup \{m\}) = \sum_{\mathcal{T} \subseteq \mathcal{V}} \boldsymbol{p}_{\mathcal{T}\cup\{m\}}. \end{aligned} \tag{26}$$

This proves the marginal decomposition. □

A.2 Proof of Proposition 2

Proposition 2 (Subset-restricted dependence)**.** *For fixed model parameters and any non-empty $\mathcal{U} \subseteq [M]$, $H_{\boldsymbol{x}}(\mathcal{U})$ depends only on $(\boldsymbol{x}_m)_{m\in\mathcal{U}}$.*

Proof. For fixed model parameters, write $\boldsymbol{h}_{\mathcal{U}}(\boldsymbol{x})$ for the state constructed from modality representations $\boldsymbol{x} = (\boldsymbol{x}_m)_{m\in[M]}$. We prove by induction on $|\mathcal{U}|$ that, for any non-empty $\mathcal{U} \subseteq [M]$ and any $\boldsymbol{x}, \boldsymbol{x}'$, if $\boldsymbol{x}_m = \boldsymbol{x}'_m$ for every $m \in \mathcal{U}$, then $\boldsymbol{h}_{\mathcal{U}}(\boldsymbol{x}) = \boldsymbol{h}_{\mathcal{U}}(\boldsymbol{x}')$.

For the base case, let $\mathcal{U} = \{m\}$. The refinement on the edge $\varnothing \xrightarrow{+m} \{m\}$ starts from the fixed boundary state $\boldsymbol{h}_{\varnothing} = \mathbf{0}$ and uses only $\boldsymbol{x}_m$ as the added-modality representation. If $\boldsymbol{x}_m = \boldsymbol{x}'_m$, the two refinements therefore have the same initial state and the same added-modality representation. An induction over the refinement steps $t = 0, \ldots, T$ gives $\boldsymbol{z}^{(t)}(\boldsymbol{x}) = \boldsymbol{z}^{(t)}(\boldsymbol{x}')$ for every t, and hence

$$\boldsymbol{h}_{\{m\}}(\boldsymbol{x}) = \boldsymbol{z}_{\varnothing\to\{m\}}(\boldsymbol{x}) = \boldsymbol{z}_{\varnothing\to\{m\}}(\boldsymbol{x}') = \boldsymbol{h}_{\{m\}}(\boldsymbol{x}'). \tag{27}$$

Now let $r \geqslant 2$ and assume the claim holds for every non-empty subset of cardinality smaller than r. Fix $\mathcal{U} \subseteq [M]$ with $|\mathcal{U}| = r$, and suppose that $\boldsymbol{x}_m = \boldsymbol{x}'_m$ for every $m \in \mathcal{U}$. For each $m \in \mathcal{U}$, define $\mathcal{V}_m \coloneqq \mathcal{U} \setminus \{m\}$. Since $|\mathcal{V}_m| = r - 1$ and $\boldsymbol{x}_j = \boldsymbol{x}'_j$ for every $j \in \mathcal{V}_m$, the induction hypothesis gives

$$\boldsymbol{h}_{\mathcal{V}_m}(\boldsymbol{x}) = \boldsymbol{h}_{\mathcal{V}_m}(\boldsymbol{x}'). \tag{28}$$

The edge $\mathcal{V}_m \xrightarrow{+m} \mathcal{U}$ starts from this predecessor state and uses $\boldsymbol{x}_m$ as its added-modality representation. Because the predecessor states agree and $\boldsymbol{x}_m = \boldsymbol{x}'_m$, induction over the refinement steps gives

$$\boldsymbol{z}_{\mathcal{V}_m \to \mathcal{U}}(\boldsymbol{x}) = \boldsymbol{z}_{\mathcal{V}_m \to \mathcal{U}}(\boldsymbol{x}') \tag{29}$$

for every $m \in \mathcal{U}$. Applying the subset-state recurrence then yields

$$\boldsymbol{h}_{\mathcal{U}}(\boldsymbol{x}) = \frac{1}{|\mathcal{U}|} \sum_{m \in \mathcal{U}} \boldsymbol{z}_{\mathcal{U}\setminus\{m\} \to \mathcal{U}}(\boldsymbol{x}) = \frac{1}{|\mathcal{U}|} \sum_{m \in \mathcal{U}} \boldsymbol{z}_{\mathcal{U}\setminus\{m\} \to \mathcal{U}}(\boldsymbol{x}') = \boldsymbol{h}_{\mathcal{U}}(\boldsymbol{x}'). \tag{30}$$

Thus $\boldsymbol{h}_{\mathcal{U}}$ is unchanged by any change to $(\boldsymbol{x}_m)_{m \notin \mathcal{U}}$, so $H_{\boldsymbol{x}}(\mathcal{U}) = \boldsymbol{h}_{\mathcal{U}}$ depends only on $(\boldsymbol{x}_m)_{m \in \mathcal{U}}$. □

B Implementation Details

Configuration of the intra-modality tower. For all experiments, we instantiate each Tower_m as a stack of L_m Pre-LN Transformer encoder blocks (Vaswani et al., 2017; Xiong et al., 2020). Each block applies multi-head self-attention followed by a positionwise two-layer feedforward network, with layer normalisation before each sublayer and a residual connection around each sublayer. The tower preserves the token dimension and feature width, so $\mathrm{Tower}_m : \mathbb{R}^{d_x \times N} \to \mathbb{R}^{d_x \times N}$. When $L_m = 0$, we define Tower_m as the identity map.

Subset predictor. The subset predictor uses the same global or positionwise output structure as the task predictor in §3.4, with layer normalisation applied to each subset state before prediction. Its parameters are shared across all non-empty modality subsets and are separate from the parameters of the final task predictor. For global tasks, we average the normalised subset state over positions and map the resulting vector to the task output space:

$$\boldsymbol{o}_{\mathcal{U}}^{\mathrm{subset}} = \boldsymbol{W}^{\mathrm{subset}} \left(\frac{1}{N} \sum_{n=1}^{N} \mathrm{LN}(\boldsymbol{h}_{\mathcal{U}}[n]) \right) + \boldsymbol{b}^{\mathrm{subset}}. \tag{31}$$

For positionwise tasks, we apply the same affine map independently at each position:

$$\boldsymbol{o}_{\mathcal{U}}^{\mathrm{subset}}[n] = \boldsymbol{W}^{\mathrm{subset}} \mathrm{LN}(\boldsymbol{h}_{\mathcal{U}}[n]) + \boldsymbol{b}^{\mathrm{subset}}. \tag{32}$$

The same $\boldsymbol{W}^{\mathrm{subset}}$ and $\boldsymbol{b}^{\mathrm{subset}}$ therefore define $\mathrm{Pred}_{\mathrm{subset}}$ for every non-empty $\mathcal{U} \subseteq [M]$.

Task loss and fixed target prior. We choose ℓ_{task} according to the prediction type and use the same loss definition for the final task output, every subset output, and the fixed target prior in §3.5. For global multiclass classification, ℓ_{task} is the cross-entropy loss computed from the predicted logits. For scalar regression, we use the L_1 loss. For positionwise semantic segmentation, we average the cross-entropy loss over the valid positions of each example:

$$\ell_{\mathrm{task}}(\boldsymbol{o}, Y) := \frac{1}{|\mathcal{I}|} \sum_{n \in \mathcal{I}} \mathrm{CE}(\boldsymbol{o}[n], Y[n]), \tag{33}$$

where $\mathcal{I} \subseteq [N]$ denotes the set of valid positions. Each loss is defined per training example before taking the expectations in $\mathcal{L}_{\mathrm{main}}$ and $\mathcal{L}_{\mathrm{subset}}$.

We estimate the fixed input-independent target prior once from the training targets and keep it unchanged during optimisation. For global multiclass classification, the prior predicts the empirical class distribution of the training set. Writing q_c for the empirical frequency of class c, its loss for target Y is $\ell_{\mathrm{prior}}(Y) = -\log q_Y$. For scalar regression with L_1 loss, $\ell_{\mathrm{prior}}(Y) = |\widetilde{Y}_{\mathrm{train}} - Y|$, where $\widetilde{Y}_{\mathrm{train}}$ is the median training target. For positionwise semantic segmentation, the prior predicts at every valid position the empirical class distribution computed over all valid training positions. If q_c^{seg} denotes this distribution, the corresponding prior loss is

$$\ell_{\mathrm{prior}}(Y) = \frac{1}{|\mathcal{I}|} \sum_{n \in \mathcal{I}} \left(-\log q^{\mathrm{seg}}_{Y[n]} \right). \tag{34}$$

These prior losses provide the input-independent reference used to construct the predictability proxy in §3.5.

C Extended Related Work

This appendix extends the discussion in §4.

Structured fusion and information flow. Tensor Fusion Network takes outer products of constant-augmented modality embeddings to expose unimodal, bimodal, and trimodal terms, while Low-Rank Multimodal Fusion computes tensor fusion through modality-specific factors (Zadeh et al., 2017; Liu et al., 2018). Dynamic Fusion Graph combines subset representations through node-specific networks and sample-dependent edge efficacies computed jointly from unimodal inputs (Zadeh et al., 2018a). Factorized Multimodal Transformer assigns attention factors to modality combinations and merges their outputs through learned summarisation networks (Zadeh et al., 2019). Multimodal Residual Optimization sums subset-specific predictions and blocks gradients through lower-order terms, with simultaneous optimisation or sequential freezing of previously trained orders (Wörtwein et al., 2022). Deep Sets pools element embeddings, while Set Transformer models dependencies through attention, both targeting permutation-invariant set representations (Zaheer et al., 2017; Lee et al., 2019). Non-linear transformations after pooling can introduce interactions, so permutation invariance alone leaves interaction order unspecified. Zorro masks attention to preserve unimodal representations alongside fused representations (Recasens et al., 2023). LatFuse organises fusion into subset-restricted states of identified modalities and constructs them through a shared operator on Boolean-lattice covering edges (§3.3).

Context-conditioned attention and gating. MulT models dependencies between unaligned sequences through directional cross-modal attention, and Perceiver repeatedly aggregates inputs into a latent bottleneck through asymmetric attention (Tsai et al., 2019; Jaegle et al., 2021). Gated Multimodal Units regulate modality contributions to hidden activations through multiplicative gates (Ovalle et al., 2017). Quality-aware Multimodal Fusion uses uncertainty estimates to guide dynamic fusion under varying input quality (Zhang et al., 2023c). LatFuse conditions each single-modality extension on a predecessor subset and uses a null attention slot to regulate incoming content (§3.2).

Interaction measurement and learning. Empirical Multimodally-Additive Projection assesses interactions through an additive approximation of predictions; its image-text experiments show that task accuracy can conceal limited reliance on non-additive interactions (Hessel & Lee, 2020). MM-SHAP measures modality contributions through Shapley attribution, and InterSHAP quantifies cross-modal interactions through Shapley interaction values (Parcalabescu & Frank, 2023; Wenderoth et al., 2025). Partial Information Decomposition (PID) (Williams & Beer, 2010) distinguishes target-relevant information shared across sources, specific to one source, or available only jointly, termed redundancy, uniqueness, and synergy, respectively (Liang et al., 2023). Yang et al. (2025a) estimate these interaction types at the sample level, while Amit et al. (2025) analyse predictive information in internal multimodal embeddings. MMoE learns experts for different interaction types and gates their outputs (Yu et al., 2024). I^2MoE introduces weakly supervised interaction losses and reweights expert contributions for sample-level and dataset-level interpretation (Xin et al., 2025). Decomposition-based Multimodal Interaction Learning uses variational modules to separate redundant, unique, and synergistic representation components (Yang et al., 2026). The representation-valued components of LatFuse organise interactions by subset cardinality; interpreting their higher-order deviations from additivity as PID information types requires additional distributional analysis (§3.4).

Set functions and interaction decompositions. Möbius inversion decomposes set functions into subset-indexed components, and discrete lattice derivatives characterise interactions through changes in marginal contributions (Rota, 1964; Grabisch & Labreuche, 2007). For a fixed zero-boundary function, these components reconstruct every subset value exactly, and the absence of components above first order is equivalent to additivity and context-independent marginal contributions (Rota, 1964; Grabisch & Labreuche, 2007). The Shapley–Taylor interaction index extends axiomatic Shapley attribution to feature subsets up to a selected order (Sundararajan et al., 2020). Under a product measure, functional analysis of variance (fANOVA) centres each component with respect to every participating variable, yielding orthogonal functions and additive variance contributions (Sobol', 2001). In LatFuse, exact reconstruction concerns subset states before non-linear readout, and identifying the components with fANOVA terms would require distributional centring beyond samplewise cosine regularisation (§§3.4 and 3.5).

Learning and inference with missing modalities. ModDrop randomly omits input channels during training, while HeMIS pools the mean and variance of available modality embeddings for segmentation (Neverova et al., 2016; Havaei et al., 2016). Everything at Once combines a shared fusion transformer with combinatorial contrastive learning to support different modality combinations in video retrieval (Shvetsova et al., 2022). Flex-MoE couples a learned missing-modality representation bank with availability-aware expert routing (Yun et al., 2024). CMNeXt fuses a variable set of auxiliary modalities with RGB through its Self-Query Hub and Parallel Pooling Mixer (Zhang et al., 2023b). For latent-variable inference and multimodal generation, the Multimodal Variational Autoencoder combines observed-modality posterior information through a product-of-experts inference network (Wu & Goodman, 2018). The Generalized Multimodal ELBO combines products of experts associated with modality subsets (Sutter et al., 2021). For fixed parameters, LATFUSE preserves all subset states contained in the observed-modality sublattice, supporting discriminative inference across subsets of its fixed modality collection (Proposition 2).

Modality imbalance and optimisation. Wang et al. (2020) identify branch-specific differences in overfitting and generalisation and propose Gradient Blending to combine training signals. Under a specified late-fusion network and data model, Huang et al. (2022) show how joint optimisation can favour learning only some modalities. On-the-fly Gradient Modulation adjusts modality-specific gradients according to their contributions to the training objective (Peng et al., 2022). Wu et al. (2022) measure conditional modality utilisation through the accuracy gain from adding a modality and use it to balance learning speeds. LATFUSE supervises every non-empty subset through a shared predictor with equal weighting across orders and calibrates edge gates from gains in a monotone predictability proxy (§3.5).

D EXPERIMENTAL DETAILS

Dataset and evaluation details. Table 8 summarises the feature sources, input dimensions, evaluated samples, data splits, and reporting protocols for the four benchmarks.

Table 8: Dataset and evaluation details for the four benchmarks.

Dataset	Item	Configuration
TCGA-BRCA	Feature source	MOGONET preprocessed features
	Feature dimensions	mRNA: 1000, miRNA: 503, DNA: 1000
	Evaluated data	875 matched patients
	Data split	Stratified five-fold cross-validation
	Reporting	Fold mean within each seed, mean $\pm$ std across 20 seeds
MLOmics-Pan-cancer	Feature source	`Original` features
	Feature dimensions	mRNA: 3217, miRNA: 383, DNA: 3139, CNV: 3105
	Evaluated data	8,314 patients
	Data split	Stratified five-fold cross-validation
	Reporting	Fold mean within each seed, mean $\pm$ std across 5 seeds
CMU-MOSI	Feature source	Raw words, COVAREP acoustic, Facet visual
	Feature dimensions	Language: 768, acoustic: 5, visual: 20
	Evaluated data	2,199 utterances
	Data split	1,284 / 229 / 686 for train / validation / test
	Reporting	Test mean $\pm$ std across 5 seeds
DeLiVER	Feature source	RGB, Depth, Event, LiDAR frames
	Evaluated data	7,885 front-view samples
	Data split	3,983 / 2,005 / 1,897 for train / validation / test
	Reporting	Validation mIoU mean $\pm$ std across 3 seeds

For TCGA-BRCA, we use the preprocessed cohort distributed with MOGONET and retain patients with matched mRNA expression, DNA methylation, miRNA expression, and PAM50 subtype labels (Wang et al., 2021). For MLOmics-Pan-cancer, we use the `Original` feature version containing the four omics types for 32 cancer classes (Yang et al., 2025b). For both multiomics benchmarks, we compute each metric on every held-out fold, average the five fold scores within each seed, and report the mean and standard deviation across seeds. On MLOmics-Pan-cancer, PREC denotes multiclass

precision weighted by class support, while NMI and ARI measure agreement between the predicted and ground-truth class assignments.

CMU-MOSI follows the official split by source video, which keeps all segments from the same video within one split (Zadeh et al., 2016; 2018b). For ACC_7, we clip continuous predictions to $[-3, 3]$ and round the predictions and targets to the nearest integer sentiment level. For the binary evaluations, *has0* separates negative scores from non-negative scores, while *non0* removes zero targets and separates negative scores from positive scores. The corresponding binary F_1 scores use support-weighted averaging. MAE and Pearson correlation are computed directly from the continuous sentiment predictions and targets.

For DeLiVER, we follow the front-view semantic segmentation protocol with 25 classes (Zhang et al., 2023b). Following the StitchFusion evaluation setting used for the reported baseline, we report mIoU on the official validation split (Li et al., 2025).

Data processing and training configuration. For TCGA-BRCA, we retain the feature scaling supplied with the preprocessed MOGONET features (Wang et al., 2021). For MLOmics-Pan-cancer, we retain the supplied z-score normalisation of the `Original` feature set without additional feature selection (Yang et al., 2025b). Each omics modality initially forms one token from its complete feature vector. We apply a Perceiver-style resampler with four learned latent queries to produce four modality tokens before fusion (Jaegle et al., 2021).

For CMU-MOSI, we encode the segment transcript with fine-tuned `BERT-base-uncased` and use its contextual hidden states as the language representation (Devlin et al., 2019). We use 5-dimensional COVAREP acoustic features and 20-dimensional Facet visual features from the word-aligned CMU Multimodal SDK representation (Zadeh et al., 2018b). When BERT splits a source word into multiple WordPiece tokens, we replicate the acoustic and visual vector aligned to that word at each corresponding WordPiece position (Rahman et al., 2020). We assign zero acoustic and visual vectors to special-token and padding positions and truncate or pad each sequence to 50 positions (Rahman et al., 2020). The resulting padding mask is applied to attention operations involving the MOSI sequences.

For DeLiVER, we retain the input processing, image resolution, and data augmentation protocol of StitchFusion (Li et al., 2025). The integration of LATFUSE with the segmentation architecture is described separately below.

We run TCGA-BRCA, MLOmics-Pan-cancer, and CMU-MOSI experiments on one NVIDIA GeForce RTX 4090 and DeLiVER experiments on one NVIDIA A100. Table 9 summarises the benchmark-specific fusion configuration and objective coefficients used in the reported experiments.

Table 9: Model configuration across the evaluation benchmarks.

Configuration	TCGA-BRCA	MLOmics-Pan-cancer	CMU-MOSI	DeLiVER
Encoder or backbone	Perceiver	Perceiver	BERT-base (language)	StitchFusion
Tokens N	4	4	50	Spatial tokens
Tower_m	Identity	Identity	Self-attention	Identity
d_x	128	128	256	64
d_h	128	128	256	64
d_u	128	128	256	64
Attention neighbourhood	Global	Global	Global	Pointwise
λ_{subset}	1.0	1.0	1.0	1.0
λ_{gate}	0.2	0.2	1.0	0.0
λ_{decorr}	0.2	0.2	1.0	0.0

StitchFusion integration details. We integrate LATFUSE into the StitchFusion architecture (Li et al., 2025) by replacing the attention-side MultiAdapter injection at selected Transformer blocks. Let $\boldsymbol{x}_m^{(b)} \in \mathbb{R}^{d_b \times N_b}$ denote the tokens of modality m entering block b, and let $\boldsymbol{z}_m^{(b)}$ denote the state after the self-attention residual:

$$\boldsymbol{z}_m^{(b)} = \boldsymbol{x}_m^{(b)} + \text{DropPath}\Big(\text{Attn}_b\Big(\text{LN}_1\Big(\boldsymbol{x}_m^{(b)}\Big)\Big)\Big). \tag{35}$$

At this point, StitchFusion injects a MultiAdapter message from every source modality i into every target modality $j \neq i$ (Li et al., 2025), which can be written as

$$\boldsymbol{z}_j^{(b)} \leftarrow \boldsymbol{z}_j^{(b)} + \sum_{i \neq j} \mathrm{DropPath}\Big(\mathrm{Ada}_{i\to j}^{(b)}\Big(\mathrm{LN}_1\Big(\boldsymbol{x}_i^{(b)}\Big)\Big)\Big). \tag{36}$$

This interface contains $M(M-1)$ directed single-source messages for M modality streams. We replace these messages with M joint leave-one-out updates, one for each target modality. For target j, define its source set as

$$\mathcal{S}_j \coloneqq [M] \setminus \{j\}, \qquad |\mathcal{S}_j| = M - 1. \tag{37}$$

At each selected block, LATFUSE receives the same pre-attention normalised source tokens used by the original attention-side MultiAdapter, namely $\boldsymbol{e}_m^{(b)} \coloneqq \mathrm{LN}_1(\boldsymbol{x}_m^{(b)})$. A shared truncated lattice trunk constructs the subset states

$$\Big\{\boldsymbol{h}_{\mathcal{U}}^{(b)} : \varnothing \neq \mathcal{U} \subsetneq [M]\Big\}, \tag{38}$$

so the construction stops at order $M-1$ and reuses every overlapping subset state across the M leave-one-out source sets. For each target j, the Möbius readout is restricted to the sublattice induced by $\mathcal{S}_j$:

$$\boldsymbol{r}_{-j}^{(b)} \coloneqq \mathrm{Readout}_b\Big(\Big\{\boldsymbol{h}_{\mathcal{U}}^{(b)} : \varnothing \neq \mathcal{U} \subseteq \mathcal{S}_j\Big\}\Big) \in \mathbb{R}^{d_u \times N_b}. \tag{39}$$

The M leave-one-out evaluations share the same Möbius readout parameters within a block, while different Transformer blocks use independent LATFUSE parameters. Each leave-one-out representation is mapped to the target stream through a zero-initialised residual projection:

$$\boldsymbol{\delta}_j^{(b)} = \boldsymbol{W}_{b,j}^{\mathrm{inj}} \boldsymbol{r}_{-j}^{(b)} + \boldsymbol{b}_{b,j}^{\mathrm{inj}}, \qquad \boldsymbol{W}_{b,j}^{\mathrm{inj}} \leftarrow \mathbf{0}, \qquad \boldsymbol{b}_{b,j}^{\mathrm{inj}} \leftarrow \mathbf{0}. \tag{40}$$

We inject this residual through the original stochastic-depth path before the block MLP:

$$\boldsymbol{z}_j^{(b)} \leftarrow \boldsymbol{z}_j^{(b)} + \mathrm{DropPath}\Big(\boldsymbol{\delta}_j^{(b)}\Big). \tag{41}$$

The zero initialisation makes the added LATFUSE branch contribute zero at the start of training. For DeLiVER with $\{\mathrm{R}, \mathrm{D}, \mathrm{E}, \mathrm{L}\}$ denoting RGB, Depth, Event, and LiDAR, respectively, the four information flows are

$$\{\mathrm{D},\mathrm{E},\mathrm{L}\} \to \mathrm{R}, \qquad \{\mathrm{R},\mathrm{E},\mathrm{L}\} \to \mathrm{D}, \qquad \{\mathrm{R},\mathrm{D},\mathrm{L}\} \to \mathrm{E}, \qquad \{\mathrm{R},\mathrm{D},\mathrm{E}\} \to \mathrm{L}. \tag{42}$$

Thus, each target receives one jointly fused residual from the other $M-1$ modalities. Following the DeLiVER StitchFusion configuration (Li et al., 2025), we apply this replacement to every Transformer block in the last two MiT stages. We retain the original MLP-side MultiAdapter injection and the remaining StitchFusion encoder and decoder computations unchanged.

Ablation specifications. We use the same data splits and training protocol as the full model in §3.5.

(A1) *Removing predecessor context conditioning.* We remove the predecessor state from the input to edge refinement and retain the recursive subset construction. For each edge $\mathcal{V} \xrightarrow{+m} \mathcal{U}$, we initialise refinement at zero and add $\boldsymbol{h}_{\mathcal{V}}$ after the final refinement step.

$$\begin{aligned} \bar{\boldsymbol{z}}_m^{(0)} &\coloneqq \mathbf{0}, \qquad \bar{\boldsymbol{z}}_m^{(t+1)} \coloneqq f\Big(\bar{\boldsymbol{z}}_m^{(t)}; \boldsymbol{x}_m\Big), \quad t = 0, \dots, T-1, \\ \boldsymbol{z}_{\mathcal{V}\to\mathcal{U}}^{\mathrm{A1}} &\coloneqq \boldsymbol{h}_{\mathcal{V}} + \bar{\boldsymbol{z}}_m^{(T)}. \end{aligned} \tag{43}$$

The refinement operator and number of steps remain unchanged. We retain the predecessor averaging rule in §3.3 and the readout in §3.4.

(A2) *Removing Möbius inversion from the readout.* We remove Möbius inversion from the task readout and pass each cumulative subset state directly to its order-specific projection.

$$\boldsymbol{u}_{\mathcal{S}}^{\mathrm{A2}} \coloneqq \boldsymbol{W}_r^{\mathrm{mob}} \boldsymbol{h}_{\mathcal{S}}, \qquad \mathcal{S} \in \mathcal{G}_r. \tag{44}$$

The subset construction, order aggregation, cross-order residual fusion, and task predictor remain unchanged. We still compute $\boldsymbol{p}_{\mathcal{S}}$ by Möbius inversion for the component decorrelation loss.

(A3) *Removing identity preservation.* We remove $-\phi(\mathrm{LN}(\boldsymbol{z}^{(t)}))$ and $-\psi(\mathrm{LN}([\boldsymbol{c}_{r-1};\mathbf{0}]))$ from edge refinement and cross-order residual fusion, respectively. The recurrences use the following updates.

$$\begin{aligned}\boldsymbol{z}^{(t+1)} &\coloneqq \boldsymbol{z}^{(t)} + \boldsymbol{a}^{(t)} + \phi\Big(\mathrm{LN}\Big(\boldsymbol{z}^{(t)} + \boldsymbol{a}^{(t)}\Big)\Big)\,,\\ \boldsymbol{c}_r &\coloneqq \boldsymbol{c}_{r-1} + \boldsymbol{g}_r + \psi(\mathrm{LN}([\boldsymbol{c}_{r-1};\boldsymbol{g}_r]))\,.\end{aligned} \tag{45}$$

Both updates retain their residual additions and the remaining feedforward terms.

(A4) *Removing intermediate subset states.* We remove every subset state $\boldsymbol{h}_{\mathcal{U}}$ with $1 < |\mathcal{U}| < M$, together with its computation, supervision, and reuse. The retained unimodal states $\{\boldsymbol{h}_{\{m\}}\}_{m=1}^{M}$ connect directly to the full-set state $\boldsymbol{h}_{[M]}$ through a star topology.

(L1) *Removing the gate calibration loss.* We remove $\mathcal{L}_{\mathrm{gate}}$ from the training objective by setting $\lambda_{\mathrm{gate}} = 0$. The null attention slot and the resulting edge gates remain part of the model.

(L2) *Removing the component decorrelation loss.* We remove $\mathcal{L}_{\mathrm{decorr}}$ from the training objective by setting $\lambda_{\mathrm{decorr}} = 0$. The Möbius decomposition and readout remain unchanged.

Inference with missing modalities. All models retain their full-modality training parameters throughout evaluation. We perform no subset-specific retraining or test-time adaptation. For sample i, let $\mathcal{O}_i \subseteq [M]$ denote the observed modalities, with $|\mathcal{O}_i| = M - 1$ in every evaluated condition. Dataset-level missingness fixes one unavailable modality across the evaluation set and repeats the evaluation for each modality. For heterogeneous missingness, we uniformly shuffle a balanced list of modality indices and assign one index to each sample without using target labels. The numbers of samples assigned to any two modalities differ by at most one.

At the fusion-module level, LATFUSE computes representations for the observed modalities and constructs every non-empty subset state $\boldsymbol{h}_{\mathcal{U}}$ with $\mathcal{U} \subseteq \mathcal{O}_i$ using the recurrence in §3.3. The boundary state remains $\boldsymbol{h}_{\varnothing} = \mathbf{0}$. For fixed parameters and input representations, Proposition 2 guarantees that these states equal their counterparts in the full lattice. Möbius inversion uses only states indexed by $\mathcal{T} \subseteq \mathcal{S}$, so each retained component $\boldsymbol{p}_{\mathcal{S}}$ also equals its full-lattice counterpart.

A readout over $\mathcal{O}_i$ groups the available components by order as $\mathcal{G}_r(\mathcal{O}_i) \coloneqq \{\mathcal{S} \subseteq \mathcal{O}_i : |\mathcal{S}| = r\}$. For $r = 1, \ldots, |\mathcal{O}_i|$, we retain the trained projection $\boldsymbol{u}_{\mathcal{S}} = \boldsymbol{W}_r^{\mathrm{mob}} \boldsymbol{p}_{\mathcal{S}}$ and transformation χ and normalise by the square root of the number of available components:

$$\boldsymbol{g}_r^{(\mathcal{O}_i)} = \frac{1}{\sqrt{|\mathcal{G}_r(\mathcal{O}_i)|}} \sum_{\mathcal{S} \in \mathcal{G}_r(\mathcal{O}_i)} (\boldsymbol{u}_{\mathcal{S}} + \chi(\boldsymbol{u}_{\mathcal{S}}))\,. \tag{46}$$

Starting from $\boldsymbol{c}_0^{(\mathcal{O}_i)} = \mathbf{0}$, we apply the cross-order recurrence in §3.4 through order $|\mathcal{O}_i|$. For the multiomics classifiers, we average the final state $\boldsymbol{c}_{|\mathcal{O}_i|}^{(\mathcal{O}_i)}$ over token positions and apply the trained task predictor. The label-dependent quantities $a_{\mathcal{U}}$ and $A_{\mathcal{U}}$ serve gate calibration only during training.

For baselines with a native modality-subset interface, we provide the observed modalities through that interface. When a baseline requires a fixed input signature, we replace the unavailable preprocessed input with an all-zero tensor of the same shape.

E ADDITIONAL ABLATION RESULTS

E.1 COMPLETE ABLATION TASK METRICS

Tables 10 and 11 report the complete task metrics using the same evaluation definitions as the main experiments. Each table reports the mean and standard deviation across the same seeds used for the corresponding main ablation results.

E.2 STRUCTURAL DIAGNOSTICS OF ABLATION VARIANTS

We evaluate the four ablation variants with structural diagnostics on the same held-out evaluation splits and trained checkpoints used in the main ablation study. For each training seed, we compute each diagnostic over the complete evaluation split and average the resulting statistics across seeds. When a diagnostic requires one vector per sample, we average tokenwise states over valid positions using the same positional mask as the task output.

Table 10: Complete ablation task metrics on the two multiomics benchmarks.

	TCGA-BRCA			MLOmics-Pan-cancer					
Variant	ACC (%) ↑	w.F_1 (%) ↑	m.F_1 (%) ↑	ACC (%) ↑	w.F_1 (%) ↑	m.F_1 (%) ↑	PREC (%) ↑	NMI (%) ↑	ARI (%) ↑
Full	**89.39±0.33**	**89.75±0.42**	**87.08±0.51**	**97.71±0.08**	**97.56±0.09**	**95.26±0.17**	**97.60±0.09**	**97.37±0.08**	**96.83±0.13**
A1	85.97±0.46	86.38±0.58	83.60±0.68	95.76±0.05	95.42±0.09	92.59±0.14	96.07±0.11	95.52±0.10	94.82±0.12
A2	88.15±0.70	88.57±0.78	85.65±0.84	96.68±0.08	96.27±0.06	93.37±0.09	96.66±0.07	96.28±0.06	95.69±0.09
A3	87.17±0.85	87.56±0.91	84.64±0.95	95.89±0.10	95.66±0.13	93.14±0.19	96.18±0.12	96.16±0.09	95.51±0.14
A4	84.60±0.52	85.04±0.55	81.75±0.68	94.70±0.04	94.19±0.09	91.18±0.11	94.80±0.08	94.94±0.09	94.23±0.10
L1	86.80±0.43	87.21±0.48	84.17±0.50	95.78±0.06	95.54±0.07	92.79±0.09	96.15±0.09	96.14±0.10	95.46±0.15
L2	88.32±0.34	88.76±0.41	85.83±0.45	96.50±0.04	96.16±0.04	93.23±0.10	96.51±0.07	96.23±0.07	95.56±0.08

Table 11: Complete ablation task metrics on CMU-MOSI.

Variant	MAE ↓	Corr ↑	ACC_7 (%) ↑	ACC_2 (*has0* / *non0*) (%) ↑	F_1 (*has0* / *non0*) (%) ↑
Full	**0.662±0.007**	**0.823±0.005**	**50.95±1.30**	**86.17±0.70 / 88.66±0.58**	**86.15±0.65 / 88.61±0.57**
A1	0.709±0.011	0.795±0.005	45.42±2.02	82.76±0.95 / 84.94±0.76	82.83±0.93 / 84.91±0.76
A2	0.674±0.005	0.817±0.004	48.61±2.24	84.76±1.00 / 87.48±0.60	84.81±0.94 / 87.44±0.57
A3	0.689±0.006	0.807±0.004	47.03±1.85	84.49±0.57 / 86.67±0.51	84.57±0.57 / 86.64±0.53
A4	0.710±0.008	0.792±0.005	44.30±1.99	82.32±0.37 / 84.30±0.25	82.40±0.38 / 84.28±0.27
L1	0.680±0.004	0.810±0.005	47.93±1.06	84.55±0.49 / 86.83±0.27	84.68±0.44 / 86.82±0.23
L2	0.671±0.005	0.819±0.003	49.11±1.31	85.03±0.85 / 87.76±0.58	85.13±0.72 / 87.80±0.49

A1. We quantify cross-modal interactions with global InterSHAP, which applies the Shapley Interaction Index to modality coalitions and measures the fraction of model behaviour attributed to interactions (Wenderoth et al., 2025). For a sample a, modalities i and j, and a context subset $\mathcal{S} \subseteq [M] \setminus \{i, j\}$, we compute

$$\Delta_{ij}^{(a)}(\mathcal{S}) = v^{(a)}(\mathcal{S} \cup \{i, j\}) - v^{(a)}(\mathcal{S} \cup \{i\}) - v^{(a)}(\mathcal{S} \cup \{j\}) + v^{(a)}(\mathcal{S}), \tag{47}$$

where $v^{(a)}(\mathcal{S})$ denotes one scalar output component obtained when only modalities in $\mathcal{S}$ are available. For $\mathcal{S} = \varnothing$, we use the fixed target prior predictor as the coalition output. The Shapley interaction between modalities i and j is

$$\phi_{ij}^{(a)} = \sum_{\mathcal{S} \subseteq [M] \setminus \{i,j\}} \frac{|\mathcal{S}|!(M - |\mathcal{S}| - 2)!}{2(M-1)!} \Delta_{ij}^{(a)}(\mathcal{S}). \tag{48}$$

We enumerate all modality coalitions and follow the global InterSHAP aggregation over samples and model outputs. For classification, we use final class probabilities, and for regression, we use standardised scalar predictions. We use the native subset inference path for every non-empty coalition. Table 12 reports the resulting global InterSHAP scores. The consistently higher InterSHAP scores of the full model indicate stronger context-dependent cross-modal interactions in its output.

A2. We compare the representations supplied to the cross-order residual readout using linear centred kernel alignment (Kornblith et al., 2019). For interaction order r, we collect the pooled representations $\boldsymbol{g}_r$ from all n evaluation samples into $\boldsymbol{G}_r \in \mathbb{R}^{n \times d}$. For two orders r and s, we compute

$$\mathrm{CKA}(\boldsymbol{G}_r, \boldsymbol{G}_s) = \frac{\left\|\boldsymbol{G}_r^\top \boldsymbol{H} \boldsymbol{G}_s\right\|_{\mathrm{F}}^2}{\left\|\boldsymbol{G}_r^\top \boldsymbol{H} \boldsymbol{G}_r\right\|_{\mathrm{F}} \left\|\boldsymbol{G}_s^\top \boldsymbol{H} \boldsymbol{G}_s\right\|_{\mathrm{F}}}, \qquad \boldsymbol{H} = \boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top. \tag{49}$$

We compute the complete $M \times M$ CKA matrix for each seed and average the matrices across seeds. The full model and A2 use the same component decorrelation setting so that the comparison isolates the representations supplied to the orderwise readout. Figure 3 shows the resulting matrices on a common scale from zero to one. The lower off-diagonal CKA of the full model indicates reduced representational overlap across interaction orders after Möbius decomposition.

A3. We measure whether the edge refinement and cross-order residual update preserve their input when the corresponding added signal is zero. For the edge refinement diagnostic, we use covering edges with non-empty predecessor states from real evaluation samples and force all attention mass to the null slot at every refinement step, which gives $\boldsymbol{a}^{(t)} = \mathbf{0}$. After T refinement steps, we compute

$$E_{\mathrm{edge}} = \frac{\left\|\boldsymbol{z}_{\mathrm{null}}^{(T)} - \boldsymbol{h}_{\mathcal{V}}\right\|_{\mathrm{F}}}{\|\boldsymbol{h}_{\mathcal{V}}\|_{\mathrm{F}} + \varepsilon}. \tag{50}$$

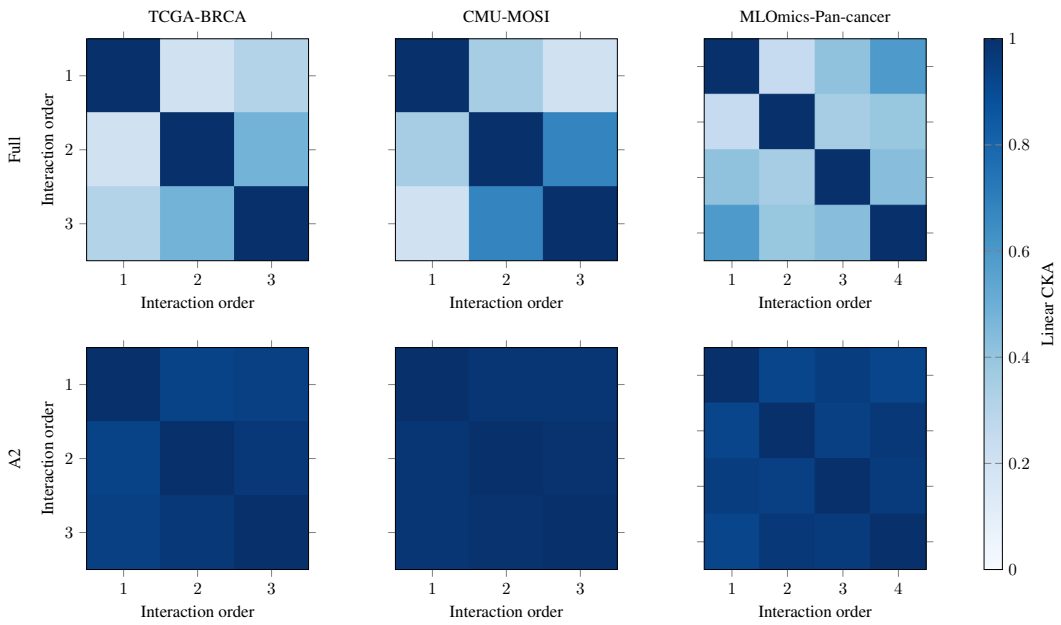


Figure 3: Orderwise representation similarity for the full model and A2 across TCGA-BRCA, CMU-MOSI, and MLOmics-Pan-cancer. The full model shows substantially lower off-diagonal CKA across all datasets, while A2 retains highly overlapping representations across interaction orders.

For the cross-order diagnostic, we use the states $\boldsymbol{c}_{r-1}$ produced by real evaluation samples for $r = 2, \ldots, M$, set $\boldsymbol{g}_r = \mathbf{0}$, and compute

$$E_{\text{order}} = \frac{\|\boldsymbol{c}_r(\boldsymbol{g}_r = \mathbf{0}) - \boldsymbol{c}_{r-1}\|_{\text{F}}}{\|\boldsymbol{c}_{r-1}\|_{\text{F}} + \varepsilon}. \tag{51}$$

For each seed, we take the median error across all applicable samples and edges or interaction orders. Table 12 reports both relative identity errors. The full model maintains both errors near numerical zero, while A3 produces non-zero state drift in the edge refinement and cross-order integration.

A4. We use linear CKA to measure the representation similarity between the states that directly enter the full-modality node and the resulting full-modality state. For the full lattice, we form one evaluation matrix $\boldsymbol{H}_{-m}$ from each direct predecessor $\boldsymbol{h}_{[M]\setminus\{m\}}$ and form $\boldsymbol{H}_{\text{full}}$ from $\boldsymbol{h}_{[M]}$. We compute $\text{CKA}(\boldsymbol{H}_{-m}, \boldsymbol{H}_{\text{full}})$ for every modality m and average the resulting values across modalities. For A4, the star topology directly connects each singleton state to the full-modality node, so we compute $\text{CKA}(\boldsymbol{H}_m^{\text{A4}}, \boldsymbol{H}_{\text{full}}^{\text{A4}})$ for each source modality and average across modalities. Table 12 reports the resulting source-to-full CKA values. The higher source-to-full CKA of the full lattice indicates greater representational continuity when intermediate subset states precede the full-modality state.

Table 12: Structural diagnostics for A1, A3, and A4.

	A1 (InterSHAP)		**A3** ($E_{\text{edge}}/E_{\text{order}}$)		**A4** (CKA)	
Dataset	Full	A1	Full	A3	Full	A4
TCGA-BRCA	0.207	0.097	≈ 0 / ≈ 0	0.898 / 1.039	0.965	0.471
CMU-MOSI	0.210	0.089	≈ 0 / ≈ 0	0.143 / 0.320	0.676	0.343
MLOmics-Pan-cancer	0.177	0.074	≈ 0 / ≈ 0	0.924 / 0.803	0.752	0.221

F Complete Missing-Modality Evaluation

Table 13 reports the complete metric set for LatFuse and two representative baselines per benchmark under the same dataset-level and heterogeneous conditions used in Table 7.

Table 13: Complete task metrics for the missing-modality protocols used in Table 7.

TCGA-BRCA	Metric	Method	Full	−mRNA	−miRNA	−DNA		Heterogeneous
	ACC	MM-Dynamics	86.94±0.78	53.95±10.48	81.03±3.28	83.59±1.18		72.96±4.59
		MOHGCN	87.91±0.38	77.47±1.37	85.17±1.72	84.81±1.70		82.50±1.77
		LatFuse	**89.39±0.33**	**79.99±0.84**	**88.48±0.45**	**86.83±0.74**		**85.39±0.48**
	F_1 weighted	MM-Dynamics	87.33±0.83	51.71±13.38	79.73±4.29	83.57±1.35		73.99±4.29
		MOHGCN	88.25±0.39	77.67±1.50	85.37±1.93	85.07±1.91		82.76±1.85
		LatFuse	**89.75±0.42**	**79.64±0.93**	**88.84±0.46**	**87.20±0.78**		**85.59±0.52**
	F_1 macro	MM-Dynamics	83.98±1.16	51.32±7.43	76.84±4.59	79.20±1.97		71.05±3.61
		MOHGCN	85.39±0.84	73.07±2.12	82.57±2.56	81.74±2.02		79.20±2.36
		LatFuse	**87.08±0.51**	**76.38±1.41**	**85.42±0.70**	**83.92±0.95**		**82.77±0.79**
MLOmics-Pan-cancer	**Metric**	**Method**	Full	−mRNA	−miRNA	−DNA	−CNV	Heterogeneous
	ACC	DMIB	96.89±0.19	95.69±0.22	96.43±0.07	96.21±0.10	96.79±0.08	96.36±0.04
		MOSGAT	96.67±0.31	94.74±0.29	95.71±0.13	95.92±0.37	96.68±0.11	95.69±0.19
		LatFuse	**97.71±0.08**	**96.04±0.15**	**96.54±0.05**	**96.25±0.10**	**96.91±0.09**	**96.63±0.13**
	F_1 weighted	DMIB	96.80±0.19	95.45±0.26	96.30±0.05	95.92±0.14	96.67±0.09	96.20±0.04
		MOSGAT	96.53±0.37	94.77±0.25	95.49±0.10	95.79±0.39	96.52±0.11	95.58±0.19
		LatFuse	**97.56±0.09**	**95.83±0.20**	**96.42±0.07**	**96.26±0.13**	**96.81±0.07**	**96.37±0.17**
	F_1 macro	DMIB	94.57±0.67	91.70±0.82	93.58±0.22	91.84±0.68	93.30±0.39	92.85±0.41
		MOSGAT	92.00±1.11	87.54±0.86	87.86±0.53	89.28±1.23	91.05±0.56	88.83±0.76
		LatFuse	**95.26±0.17**	**92.49±0.33**	**93.74±0.17**	**92.59±0.26**	**93.79±0.18**	**93.58±0.17**
	PREC	DMIB	96.92±0.19	95.81±0.23	96.53±0.06	96.19±0.20	96.82±0.10	96.43±0.07
		MOSGAT	96.75±0.34	95.45±0.16	95.75±0.06	96.09±0.38	96.65±0.11	95.90±0.16
		LatFuse	**97.60±0.09**	**95.97±0.17**	**96.98±0.06**	**96.34±0.09**	**97.35±0.09**	**97.07±0.14**
	NMI	DMIB	96.09±0.23	95.25±0.18	95.93±0.06	95.94±0.11	96.28±0.06	95.87±0.03
		MOSGAT	96.12±0.28	94.14±0.31	95.44±0.15	95.53±0.27	96.31±0.13	95.23±0.19
		LatFuse	**97.37±0.08**	**95.54±0.15**	**96.64±0.07**	**96.00±0.11**	**97.01±0.10**	**96.74±0.12**
	ARI	DMIB	95.20±0.29	93.79±0.29	94.85±0.12	94.68±0.09	95.28±0.05	94.73±0.06
		MOSGAT	95.48±0.35	93.04±0.30	94.62±0.12	94.65±0.32	95.54±0.09	94.38±0.17
		LatFuse	**96.83±0.13**	**94.40±0.17**	**95.57±0.06**	**94.91±0.16**	**96.52±0.14**	**95.64±0.20**
DeLiVER	**Metric**	**Method**	Full	−RGB	−Depth	−Event	−LiDAR	Heterogeneous
	mIoU	StitchFusion	68.21	52.52	37.08	66.89	66.80	55.11
		Mul-VMamba	68.99	57.68	33.96	**68.98**	68.62	55.71
		LatFuse	**69.72±0.30**	**60.18±2.77**	**41.88±3.10**	68.64±0.90	**68.82±0.77**	**59.63±1.15**

G Synthetic Order-Resolved Diagnostics

This appendix provides the target functions, diagnostic definitions, numerical evaluation procedure, and detailed results for the synthetic analysis in §5.5.

G.1 Experimental Setup

We evaluate LatFuse on two three-modality regression settings with $M = 3$ and $N = 1$. Each modality provides one scalar input, sampled independently as

$$s_1, s_2, s_3 \overset{\text{i.i.d.}}{\sim} \text{Uniform}[-\sqrt{3}, \sqrt{3}]. \tag{52}$$

Each input has zero mean and unit variance. We construct the targets using $\rho(s) := \frac{\sqrt{5}}{2}(s^2 - 1)$ and the following three functions:

$$B_1(\boldsymbol{s}) := \frac{\rho(s_1) + \rho(s_2) + \rho(s_3)}{\sqrt{3}}, \qquad B_2(\boldsymbol{s}) := s_1 s_2, \qquad B_3(\boldsymbol{s}) := s_1 s_2 s_3. \tag{53}$$

Under the input distribution, B_1, B_2, and B_3 have zero mean, unit variance, and pairwise zero covariance.

The non-linear additive setting uses $Y_{\text{add}} := B_1(\boldsymbol{s})$ with order-specific reference functions

$$G_1(\boldsymbol{s}) := B_1(\boldsymbol{s}), \qquad G_2(\boldsymbol{s}) := 0, \qquad G_3(\boldsymbol{s}) := 0. \tag{54}$$

Its non-zero functional components are

$$f^{\star}_{\{m\}}(s_m) := \frac{\rho(s_m)}{\sqrt{3}}, \qquad m \in \{1, 2, 3\}. \tag{55}$$

Each non-linear term depends on a single modality, so the target contains only first-order functional effects.

The mixed-order setting uses

$$Y_{\text{mix}} := \frac{B_1(\boldsymbol{s}) + B_2(\boldsymbol{s}) + B_3(\boldsymbol{s})}{\sqrt{3}}. \tag{56}$$

Its order-specific reference functions are

$$G_r(\boldsymbol{s}) \coloneqq \frac{B_r(\boldsymbol{s})}{\sqrt{3}}, \qquad r = 1, 2, 3. \tag{57}$$

The non-zero functional components are

$$f^\star_{\{m\}}(s_m) \coloneqq \frac{\rho(s_m)}{3}, \qquad m \in \{1, 2, 3\}, \tag{58}$$

$$f^\star_{\{1,2\}}(s_1, s_2) \coloneqq \frac{s_1 s_2}{\sqrt{3}}, \qquad f^\star_{\{1,2,3\}}(\boldsymbol{s}) \coloneqq \frac{s_1 s_2 s_3}{\sqrt{3}}. \tag{59}$$

The components $f^\star_{\{1,3\}}$ and $f^\star_{\{2,3\}}$ are zero. Both targets have zero mean and unit variance, and the three reference orders in the mixed setting each contribute variance $1/3$. The additive target has Sobol indices $S^\star_1 = S^\star_2 = S^\star_3 = 1/3$, with all higher-order indices equal to zero. The mixed target has $S^\star_1 = S^\star_2 = S^\star_3 = 1/9$, $S^\star_{12} = S^\star_{123} = 1/3$, and $S^\star_{13} = S^\star_{23} = 0$.

For each setting, we independently sample 50,000 training examples, 10,000 validation examples, and 20,000 test examples. We repeat the experiments with five random seeds. Within each seed, the two settings share the sampled inputs and data splits and differ only in the target function. The model receives the raw scalars as modality inputs $\boldsymbol{e}_m = [s_m] \in \mathbb{R}^{1\times 1}$ and produces a scalar task prediction. We retain the complete architecture in §3 and use squared error for the main task and shared subset predictor. The training objective includes $\mathcal{L}_{\text{main}}$, $\mathcal{L}_{\text{subset}}$, $\mathcal{L}_{\text{gate}}$, and $\mathcal{L}_{\text{decorr}}$ as defined in §3.5. We use G_r and $f^\star_{\mathcal{S}}$ only for diagnostic evaluation after training. For each run, we select the checkpoint with the lowest validation task MSE and freeze its parameters. All diagnostics use deterministic evaluation mode.

G.2 Order-Resolved Functional Decomposition

Let $\boldsymbol{c}_0 = \mathbf{0}$ and let $\boldsymbol{c}_1, \boldsymbol{c}_2, \boldsymbol{c}_3$ denote the cross-order residual states in §3.4. The final prediction is

$$\widehat{f}(\boldsymbol{s}) = \boldsymbol{W}^{\text{out}} \boldsymbol{c}_3(\boldsymbol{s}) + \boldsymbol{b}^{\text{out}}. \tag{60}$$

Using the final output weights, we extract the prediction increment introduced at stage r:

$$d_r(\boldsymbol{s}) \coloneqq \boldsymbol{W}^{\text{out}} \left(\boldsymbol{c}_r(\boldsymbol{s}) - \boldsymbol{c}_{r-1}(\boldsymbol{s})\right), \qquad r = 1, 2, 3. \tag{61}$$

These increments satisfy $\widehat{f}(\boldsymbol{s}) = \boldsymbol{b}^{\text{out}} + \sum_{r=1}^{3} d_r(\boldsymbol{s})$. We extract them directly from the normal forward computation. The index r identifies a readout stage, whose non-linear fusion can produce effects of multiple functional orders.

We apply functional ANOVA (Sobol', 2001) to each scalar function d_r under the prescribed independent input distribution. For each stage, we define $d_{r,\varnothing} \coloneqq \mathbb{E}[d_r(\boldsymbol{s})]$. For every non-empty $\mathcal{S} \subseteq \{1, 2, 3\}$, we recursively compute

$$d_{r,\mathcal{S}}(\boldsymbol{s}_{\mathcal{S}}) \coloneqq \mathbb{E}[d_r(\boldsymbol{s}) \mid \boldsymbol{s}_{\mathcal{S}}] - \sum_{\mathcal{T} \subsetneq \mathcal{S}} d_{r,\mathcal{T}}(\boldsymbol{s}_{\mathcal{T}}). \tag{62}$$

This decomposition gives

$$d_r(\boldsymbol{s}) = d_{r,\varnothing} + \sum_{\varnothing \neq \mathcal{S} \subseteq \{1,2,3\}} d_{r,\mathcal{S}}(\boldsymbol{s}_{\mathcal{S}}). \tag{63}$$

The subset cardinality $|\mathcal{S}|$ specifies the functional order of $d_{r,\mathcal{S}}$. For each non-empty subset, we compute the variance component

$$V_{r,\mathcal{S}} \coloneqq \mathbb{E}\left[d_{r,\mathcal{S}}(\boldsymbol{s}_{\mathcal{S}})^2\right]. \tag{64}$$

The components satisfy $\text{Var}(d_r) = \sum_{\mathcal{S} \neq \varnothing} V_{r,\mathcal{S}}$. When $\text{Var}(d_r) > 0$, the corresponding Sobol index is

$$S_{\mathcal{S}}[d_r] \coloneqq \frac{V_{r,\mathcal{S}}}{\text{Var}(d_r)}. \tag{65}$$

Each index measures a subset's share of its own increment's variance. We retain the seven subsets $\{1\}$, $\{2\}$, $\{3\}$, $\{1, 2\}$, $\{1, 3\}$, $\{2, 3\}$, and $\{1, 2, 3\}$ separately. Order-level summaries sum the indices

or variance components over subsets of the same cardinality. We report absolute variances alongside the normalised indices to preserve the scale of each increment.

To assess correspondence with the reference functions, we centre each increment as $\widetilde{d}_r(\boldsymbol{s}) := d_r(\boldsymbol{s}) - \mathbb{E}[d_r(\boldsymbol{s})]$ and absorb the stage means into the output intercept. The stagewise error is

$$\mathrm{MSE}_r := \frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \left[\widetilde{d}_r(\boldsymbol{s}^{(i)}) - G_r(\boldsymbol{s}^{(i)}) \right]^2 . \tag{66}$$

Mean subtraction is the only adjustment applied to the increments before comparison with G_r. The final task MSE measures the error of $\widehat{f}$ against the target on the same test set.

For the mixed setting, we inspect d_3 using centred individual conditional expectation curves (Goldstein et al., 2013). We use three pairs of fixed input values:

$$(s_1, s_2) \in \{(1,1), (1,-1), (1,0)\}. \tag{67}$$

For each pair, we vary $s_3 = t$ over $[-\sqrt{3}, \sqrt{3}]$. The centred response and its reference are

$$\mathrm{cICE}_{a,b}(t) := d_3(a,b,t) - d_3(a,b,0), \qquad \mathrm{cICE}^{\star}_{a,b}(t) := \frac{ab\,t}{\sqrt{3}}. \tag{68}$$

All three modalities remain observed, and the curves describe conditioning on input values. We use the same fixed evaluation grid for all runs.

G.3 Numerical Evaluation

We compute distributional expectations with tensor-product Gauss–Legendre quadrature. Let $\{(\xi_j, \lambda_j)\}_{j=1}^{Q}$ denote the one-dimensional nodes and weights on $[-1,1]$. We transform them to the input distribution using $s_j := \sqrt{3}\,\xi_j$ and $w_j := \lambda_j/2$. The probability weight of $(s_{j_1}, s_{j_2}, s_{j_3})$ is $w_{j_1} w_{j_2} w_{j_3}$. We evaluate all three increments at each grid point and marginalise the unconditioned dimensions by weighted summation. The reported decomposition uses $Q = 48$, with $Q = 32$ as a numerical convergence check. If any $V_{r,\mathcal{S}}$ changes by more than 10^{-3} between these grids, we repeat the computation with $Q = 64$. This tolerance uses the unit target variance scale.

Before evaluating the trained models, we apply the same fANOVA implementation to both analytic targets. We numerically verify functional reconstruction, zero-mean component constraints, conditional zero-mean constraints, and variance reconstruction. Task MSE and stagewise MSE use the independent test set, while distributional means and fANOVA components use quadrature. We perform the complete diagnostic procedure separately for each run before aggregating results across the five seeds. The error table reports means and standard deviations, and the variance tables and Sobol heatmap report means. We compute Sobol indices within each run before averaging, so their means can differ from ratios of the mean variance components.

G.4 Diagnostic Results

The subset-level Sobol heatmap in Figure 2 summarises the functional composition of the readout increments. We supplement it with prediction errors, absolute variance components, and conditional response curves.

Prediction and stagewise errors. Table 14 reports final task MSE and the error of each centred increment relative to its reference function.

Table 14: Prediction and stagewise functional errors.

Setting	$\mathrm{MSE}_{\mathrm{task}}$ ($\times 10^{-6}$) $\downarrow$	$\mathrm{MSE}_1 \downarrow$	$\mathrm{MSE}_2 \downarrow$	$\mathrm{MSE}_3 \downarrow$
Additive	4.530±0.369	0.02460±0.0085	0.01775±0.0059	0.00089±0.0006
Mixed	9.967±0.759	0.09873±0.0121	0.13117±0.0165	0.04382±0.0014

The final task MSEs are 4.530×10^{-6} and 9.967×10^{-6} for the additive and mixed settings, respectively. The larger stagewise errors are consistent with complementary residual corrections that combine into an accurate final prediction.

Order and subset composition. Table 15 aggregates the normalised Sobol indices and absolute fANOVA variances by interaction order. The subset-resolved variances in Table 16 retain the absolute scale of the effects shown in the main-text heatmap.

Table 15: Order composition of the readout increments, averaged over five runs. Columns labelled O_k sum the indicated quantity over subsets of cardinality k.

		Sobol indices (%)			**fANOVA variance**			
Setting	**Stage**	O_1	O_2	O_3	O_1	O_2	O_3	$\mathrm{Var}(d_r)$
Additive	d_1	**100.00**	≈ 0	≈ 0	**0.716**	≈ 0	≈ 0	0.716
	d_2	**99.83**	0.17	≈ 0	**0.018**	≈ 0	≈ 0	0.018
	d_3	**91.88**	7.99	0.13	≈ 0	≈ 0	≈ 0	≈ 0
Mixed	d_1	**100.00**	≈ 0	≈ 0	**0.168**	≈ 0	≈ 0	0.168
	d_2	35.97	**64.03**	≈ 0	0.109	**0.195**	≈ 0	0.304
	d_3	5.24	6.54	**88.22**	0.019	0.025	**0.333**	0.377

Table 16: Subset-resolved fANOVA variance components, averaged over five runs.

Setting	**Stage**	$V_{r,1}$	$V_{r,2}$	$V_{r,3}$	$V_{r,12}$	$V_{r,13}$	$V_{r,23}$	$V_{r,123}$
Additive	d_1	0.23853	0.23784	0.23985	≈ 0	≈ 0	≈ 0	≈ 0
	d_2	0.00602	0.00596	0.00570	≈ 0	≈ 0	≈ 0	≈ 0
	d_3	≈ 0	≈ 0	≈ 0	≈ 0	≈ 0	≈ 0	≈ 0
Mixed	d_1	0.05135	0.05056	0.06563	≈ 0	≈ 0	≈ 0	≈ 0
	d_2	0.04733	0.04895	0.01300	0.19274	≈ 0	≈ 0	≈ 0
	d_3	0.00463	0.00455	0.00932	0.02516	≈ 0	≈ 0	0.33302

In the additive setting, all three readout increments remain first-order dominated. The variance falls from 0.716 in d_1 to 0.018 in d_2, with a smaller contribution from d_3. The higher-order percentages in d_3 therefore describe effects within a small residual increment.

For the mixed target, second-order effects account for 64.03% of $\mathrm{Var}(d_2)$, and third-order effects account for 88.22% of $\mathrm{Var}(d_3)$. The prescribed pair $\{1, 2\}$ contributes 62.59% within d_2, while each other pair contributes less than 1% (Figure 2). The absolute three-way variance in d_3 is 0.33302, close to the reference variance $1/3$, alongside smaller lower-order components.

Conditional responses. Figure 4 shows the centred conditional responses of d_3.

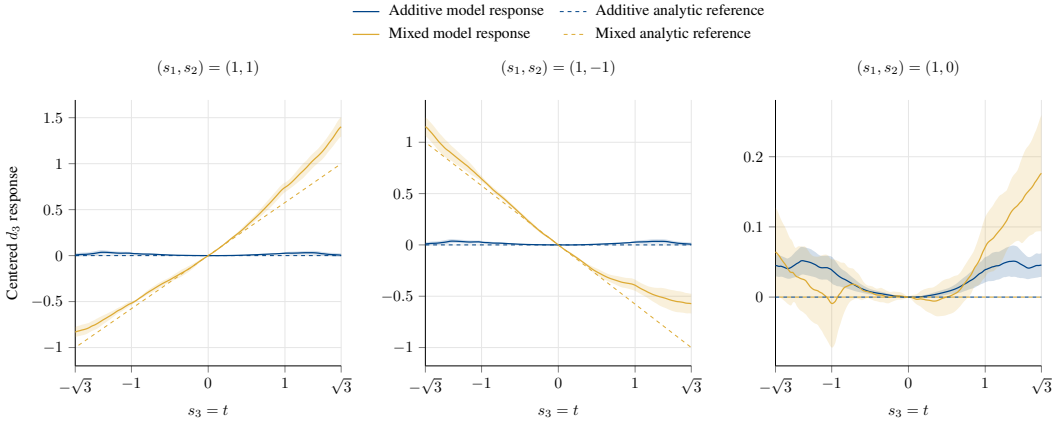


Figure 4: Centred conditional response of the third readout increment d_3 as $s_3 = t$ varies with (s_1, s_2) fixed to $(1, 1)$, $(1, -1)$, and $(1, 0)$. The analytic reference responses are $t/\sqrt{3}$, $-t/\sqrt{3}$, and 0, respectively.

The response increases with s_3 for $(s_1, s_2) = (1, 1)$, decreases for $(s_1, s_2) = (1, -1)$, and becomes strongly attenuated for $(s_1, s_2) = (1, 0)$. Across the evaluated range, the learned curves broadly follow the analytic references. All three slices fix $s_1 = 1$, where the reference functions $s_1 s_2 s_3/\sqrt{3}$ and $s_2 s_3/\sqrt{3}$ coincide. The fANOVA decomposition provides the distribution-wide evidence for the three-way interaction, while the cICE curves illustrate the conditional response of d_3.