

Label Adaptive Gating for Learning Per Label Modality Importance in Multimodal Multi Label Clinical Diagnosis

Van-Thien Tran

Thanh-Nghi Doan

Song-Toan Tran

tstoan1512@tvu.edu.vn

Research Article

Keywords: Multimodal fusion, label-adaptive gating, multi-label classification, clinical diagnosis, multimodal deep learning

Posted Date: September 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10801506/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

A real clinical diagnosis is a multi-label problem. A patient often has several findings simultaneously, and these findings are only partly correlated. Existing fusion mechanisms were designed for a single output. They decide modality importance once per sample or once per dataset; therefore, one weighting policy must serve every label at once. We propose a label-adaptive gating (LAG). The LAG maintains one learned query vector per label and allows it to cross-attend over the stacked modality embeddings. Therefore, modality importance is determined separately for each label, learned end-to-end, with no anchor modality to nominate. We compared LAG with five representative baselines: feature concatenation, learned scalar gating, confidence-gated cross-attention, confidence-gated transformer encoder, and hypernetwork-generated classifier weights, plus per-modality unimodal models. Every model used the same encoder capacity, hyperparameters, and training protocol on three independent multi-label clinical datasets: Symile-MIMIC, HAIM-MIMIC-MM, and eICU-CRD, with 6, 14, and 8 labels, respectively. Over four seeds per mechanism per dataset, with patient-level re-splitting, LAG obtained the best average rank (1.33 of 6) and the smallest average gap to each dataset's best mechanism (-0.0005 macro-AUROC). It wins on two of the three datasets. On eICU, it is second to scalar gating by a small margin, but that margin holds at every seed. The LAG alone reports the contribution of each modality to each label. These attributions appear clinically sensible when a label has sufficient data and collapse to a uniform average when it does not.

1. Introduction

Electronic health records (EHR) routinely contain a rich mixture of data. These include structured measurements (laboratory values, vital signs, and demographics), unstructured text (clinical notes and radiology reports), and medical images (chest radiographs, CT scans, and ECG signals). Multimodal fusion is the problem of combining these sources into one predictive representation, and it has become a central topic in medical machine learning research. A large toolbox now exists, from simple concatenation to attention, feature-wise conditioning, and hypernetwork-generated parameters (reviewed in Section 2.1).

Sleeman et al. [1] and Al-Zoghby et al. [2] both note that the field still lacks consistent cross-study comparison. Therefore, practitioners select a fusion mechanism by reading one paper's leaderboard, on one dataset, and one task. However, the toolbox does not contain a fusion operator built for the output structure that clinical diagnosis actually has. Each of the above mechanisms determines modality importance either globally, once for the dataset, or locally, once per sample. None of the participants decided on a diagnosis.

This practice has a specific failure mode that is frequently overlooked. Most published multimodal fusion benchmarks in clinical machine learning focus on evaluating a single binary or risk-score target, such as in-hospital mortality, a single CheXpert finding, or a single risk level. However, the differential diagnosis is more complex.

A patient in an ICU may have several concurrent findings that are only partially related, such as pulmonary edema, cardiomegaly, and pleural effusion on the same chest radiograph, along with cardiovascular, renal, and infectious issues during the same admission. Multi-label prediction is the correct approach for real diagnostic support. It also changes the requirements for a fusion mechanism, as one fused representation must now support several labels with different modality dependencies.

This per-label variation is not a conjecture. Hayat et al. [3] reported that partitioning 25 phenotype labels by acuity and adding chest radiographs to the EHR stream increased the AUROC from 0.749 to 0.800 for mixed conditions and from 0.717 to 0.745 for chronic conditions, whereas acute conditions barely moved.

Therefore, the usefulness of a modality is a property of the individual label, not the dataset. In that study, as in the fusion literature, the finding was measured after training a model that had no way of representing it. Our starting point was to make that quantity a parameter of the architecture instead of a diagnostic applied to it.

We propose label-adaptive gating (LAG), a fusion mechanism built for this requirement. The LAG uses L learned query vectors, one per label, and each of them cross-attends over the stacked modality embeddings. Therefore, the mixture of modalities used to predict a label is a separate learned quantity that can be read directly from the attention row for that label.

No modality has to be nominated as an anchor beforehand, and no label has to share its modality preference with any other label. The design commitment was deliberately narrow. LAG changes only the granularity at which modality importance is determined, from per-dataset or per-sample to per-label. The encoders, loss, and classifier head were left untouched; therefore, any gain we measure comes from that change in granularity and not from the added capacity. The cost is the $L \times d$ parameter on top of one cross-attention block (Section 3.1).

We validated the proposal against five representative baselines that span the main fusion families in the literature: concat, gate, attention, transformer, and hyperfusion (Sections 3.2–3.6). We also trained the per-modality unimodal models. Every model used the same encoder capacity, hyperparameters, and training protocol on three independent clinical resources with genuine multi-label targets: Symile-MIMIC [4] (six co-occurring CheXpert findings), HAIM-MIMIC-MM [5] (14 CheXpert findings), and eICU-CRD [6] (eight organ-system diagnosis categories).

The three differ in modality composition, scale by two orders of magnitude, and label semantics. This allows us to separate how well a mechanism generalizes from the rewards of any cohort. Section 4.1 and Table 1 provide the full specifications of the model.

Based on this, this study makes four contributions: one architectural and three empirical.

1. Label-adaptive gating (LAG), a fusion mechanism that learns the modality importance per label rather than per sample or dataset, is formulated in Section 3.1. It requires no anchor-modality

hyperparameter, adds only L query vectors over a standard cross-attention block, and is the direct architectural realization of a per-label modality dependence that prior literature reports only post hoc.

2. A matched experimental comparison of LAG against five baseline fusion mechanisms and per-modality unimodal models on three independent, genuinely multi-label clinical datasets, under identical encoder capacity, training protocol, and label-aware evaluation (macro-AUROC over all labels, with per-label decision thresholds fitted on validation only, so that every reported comparison is threshold-free).
3. A four-seed, three-dataset result in which the proposed mechanism obtains both the best average rank (1.33 of 6) and the smallest average gap to each dataset's best mechanism (~ 0.0005 macro-AUROC), establishes LAG as the most reliable multi-label default among the mechanisms tested; the same experiment places the confidence-gated transformer encoder last on all three datasets, and last at every seed on the two large enough to resolve a per-seed ordering. Where LAG is not first — elCU, against scalar gating — we report the per-seed margin rather than aggregating it away.
4. A direct inspection of LAG's learned per-label attention matrix, reporting the numeric weights and reading them against clinical expectations, establishes when the mechanism's per-label attribution is informative and when it collapses to uniform, rather than assuming that interpretability follows from architecture.

2. Related work

Multimodal fusion for clinical prediction has produced a rich inventory of mechanisms. Almost all of the inventories were designed and validated against a single output: one diagnosis, one outcome, or one mutually exclusive class. This section positions the proposed mechanism against the three literatures it inherits from: fusion architectures, clinical multimodal applications, and surveys that organize them. In each case, the granularity at which the modality importance is resolved stops one level short of the label. Section 2.3 then states what is new and what is inherited in LAG.

2.1 Multimodal fusion architectures

Early fusion concatenates the modality features before the joint classifier. Intermediate fusion learns modality-specific encoders and combines them at various depths, as in the multi-scale expert-gated fusion of computed tomography (CT) and MRI images [7]. Attention-based fusion followed the transformer [8], and cross-modal attention allows one modality to selectively attend to another. It has been applied to medical image analysis, for instance, to fuse deep retinal features with graph-based vascular descriptors for diabetic retinopathy detection [9], and to unify clinical representation learning.

IRENE [10] processes radiographs, chief complaints, and labs with intra- and intermodal attention. MeTra [11] fuses radiographs with clinical parameters for ICU survival, achieving AUROC 0.863 against 0.811 for images alone.

FiLM-style conditioning [12] and expert-gated fusion [7] allow one part of a network to modulate the other. Tensor fusion [13, 14] models cross-modal interactions multiplicatively. Mixture-of-experts [15] and hypernetworks [16] generate parameters conditioned on the input, with the latter motivating the hyperfusion baseline. The output of all these mechanisms is a single prediction head. Therefore, their gating weights, attention maps, and generated parameters are learned to serve one decision at a time. The LAG inherits its operator from this family. What changes is the query: not a modality, token, or sample, but a label.

2.2 Clinical multimodal learning and the single-label default

Clinical multimodal work is anchored on the MIMIC-CXR-JPG [17], CheXpert labeler [18], and MIMIC-IV [19]. Its usual target is a single outcome, such as in-hospital mortality, length of stay, or mutually exclusive subtypes. The reported ranking of fusion mechanisms is therefore a ranking obtained under single-label conditions, and it has largely been assumed to be transferable. MedFuse [3] is a partial exception.

It targets both mortality and 25-label phenotyping with an LSTM fusion module that tolerates missing modalities; however, it reports one design for both tasks. It does not ask whether the label structure changes, which design is preferable. HAIM [20] and eICU-CRD [21] supply pre-extracted modality embeddings and a large multi-center ICU cohort. Both are almost always used one dataset at a time; therefore, a mechanism that wins on a given cohort cannot be distinguished from one that wins in general.

Recent clinical studies have confirmed that adding a modality and conditioning mechanism are separable sources of gain. MDLA [22] predicts breast cancer molecular subtypes from paired mammography and ultrasound images. Fusion increased the four-class Matthews correlation coefficient from 0.681 and 0.729 to 0.765, and a channel-and-spatial attention block increased it further to 0.794. Therefore, the mechanism increment is approximately the same as the modality increment. It is also obtained on four mutually exclusive subtypes; therefore, the fusion step never has to reconcile competing modality preferences among labels that are positive at once.

MedFuse [3] reaches 0.770 AUROC on 25-label phenotyping, ahead of the modulation-based baselines MMTM [23] at 0.734 and DAFT [24] at 0.737 AUROC. Its more informative result is the per-label breakdown quoted in Section 1: modality usefulness is a property of an individual label. A per-label gate can represent that; a shared sample-level gate cannot.

The single-label default persists in newer works. Sepsis onset from EHR and clinical notes across more than 50,000 ICU admissions [25], liver cancer detection from MRI and CT [26], and reliability-aware fuzzy fusion for ear disease classification [27] were each carefully designed and resolved to one mutually exclusive target.

2.3 Positioning of Label-Adaptive Gating

Zhao et al. [28] argued that the conventional early/intermediate/late taxonomy no longer fits architectures in which representation learning and fusion are interlaced rather than being separable. They proposed a mechanism-first taxonomy instead: encoder–decoder, attention, graph, generative, and constraint-based methods. We adopt that view. The mechanisms compared here differ in how modality representations are combined and conditioned, not in the depth at which the combination occurs.

We also note that the label structure does not appear in the taxonomy. Two of the survey's conclusions directly impact our design. First, no family dominates across tasks; therefore, mechanism selection demands task-specific consideration. This conclusion is drawn across heterogeneous tasks, modality pairs, and training regimes, and our matched comparison retests it with the task, encoder capacity, and protocol held fixed, varying only the fusion mechanism. Second, the survey names adaptive fusion as an open direction, meaning that models learn how much each modality matters instead of assuming that all contribute equally. Label-adaptive gating takes this direction one level finer. It learns the modality importance per label rather than per sample or dataset. This is the natural granularity once the target is multi-label, and the survey does not consider it because the works it reviews have a single label to condition on.

Two key points emerge from this analysis. First, mechanism development is naturally single-label, meaning that any conclusions about behavior in a multi-label setting are speculative and not based on direct evidence. Additionally, when multi-label clinical data are available, they usually involve a single design applied to a specific group, with baseline metrics taken from different studies that vary in encoder capabilities and tuning resources. This results in a combination of architectural and implementation effects. Our approach tackles the first issue by turning MedFuse's post-hoc observation into a planned architectural strategy, which is integrated and learned throughout the process, instead of being assessed afterward.

Our experimental design addresses the second issue: the proposed mechanism and five baselines are compared under one implementation and one protocol, on one multi-label target, across three independent datasets. We want to be precise about novelty. The operator is a standard cross-attention block [8, 10, 11]. Learning, rather than fixing modality importance, is shared with expert-gated fusion [7], FiLM-style conditioning [12], and DAFT [24]. Label-wise query vectors have precedent in multi-label text classification [29] and ICD coding from clinical text [30], where per-label attention over tokens is an established method.

This approach is novel because of its integration of queries as labels, with keys and values representing modalities. Consequently, the learned quantity forms a per-label distribution across modalities that is not parameterized by any clinical fusion mechanism.

3. Proposed mechanism and baseline fusion architectures

The proposed mechanism and all five baselines share the same entry point, so the fusion step is the only thing that varies. Each dataset supplies a fixed set of modality-specific encoders, and each encoder

produces an embedding vector of the same width. The mechanisms below differ only in what they do with that set of vectors before the classifier head observes them. Symile-MIMIC uses raw inputs and trains its encoders end-to-end (1D-CNN for ECG, DenseNet121 for CXR, and MLP for labs).

HAIM and eICU use frozen pre-extracted embeddings behind matched 2-layer MLP adapters. In every case, the fusion mechanism receives M vectors of width $d = 512$; thus, no mechanism gains an advantage from a wider or deeper encoder. The proposed mechanism is presented first (Section 3.1). The five baselines against which it is tested are described in Sections 3.2–3.6, ordered by increasing structural complexity of the fusion step, and Section 3.7 lists the training settings shared by all of them.

3.1 Proposed mechanism: LAG (Label-Adaptive Gating)

Formulation. Let the M per-modality embeddings of a sample be $e_1, \dots, e_m \in \mathbb{R}^d$ ($d = 512$), stacked row-wise into $E \in \mathbb{R}^{M \times d}$, which serves as both the keys and values. The mechanism introduces a single new parameter block: L learned query vectors $q_1, \dots, q_L \in \mathbb{R}^d$, one per label ($L = 6/14/8$ for Symile-MIMIC/HAIM/eICU), initialized from $N(0, 0.02^2)$. For each label ℓ independently, scaled dot-product attention over the modality axis gives $a_\ell = \text{softmax} \left(q_\ell (EW_K)^\top / \sqrt{d} \right) \in \Delta^{M-1}$, and the label-specific context is $c_\ell = a_\ell (EW_V)$, with an attention dropout of 0.1.

Each context vector feeds a classifier head to produce the label logit, and the model is trained with the same masked, per-label weighted BCE loss as every baseline. Because a_ℓ is a probability distribution over modalities computed separately for every ℓ , the $L \times M$ matrix $A = [a_1; \dots; a_L]$ is the mechanism's explicit statement of the modality it uses for which label.

Design rationale. Two properties follow from this construction: First, the weighting policy is per label and not per the sample. The sample-level gate of Section 3.3 must commit to one modality mixture that serves every label at once; therefore, a dataset in which the chest film decides one finding and the laboratory panel decides another forces it into a compromise. The LAG has no such constraint because each query row is free to converge on its own mixture.

Second, there is no anchor modality in the proposed method. The cross-attention baselines in Sections 3.4 and 3.5 require one modality to be nominated as the query source, which is a dataset-specific choice that must be made before training and re-made whenever the modality set changes. LAG removes that choice because the queries come from the label set, which the task has already defined. This also removes a judgment call made before training, which is usually left undocumented in the literature.

Cost. The added parameter count was $L \times d$, which was 3,072 on Symile-MIMIC, 7,168 on HAIM, and 4,096 on eICU. On top of that, one standard cross-attention block is shared across labels. Therefore, LAG is among the cheapest mechanisms tested and is far cheaper than the transformer encoder (Section 3.5) or hypernetwork (Section 3.6). It computes scales linearly in the label count, and the attention is over M modality slots rather than over a token sequence, so M stays at three in all three datasets here.

Interpretability and how we test it. Averaging the attention row for label ℓ over a test set provides a direct per-label attribution over modalities. It is read off the trained model itself and is not produced by a post-hoc attribution method. We do not treat this as guaranteed by the architecture. Section 5.3 inspects the learned matrix on all three datasets and reports where it recovers a clinically sensible structure and where it collapses to a uniform average.

3.2 Baseline: Concat

The simplest possible baseline: the per-modality embeddings are concatenated into one long vector and passed to an MLP classifier. No modality is weighted, and no interaction is modelled beyond what the head can learn from the joint vector. Concat was included because it is the honest floor of the comparison.

It has no fusion-specific parameters; therefore, any mechanism that fails to beat it does not earn its added complexity. It also has a property that more elaborate mechanisms lack: because it never discards or down-weights a modality, it cannot be misled by a poor weighting decision. This directly affects the HAIM results in Section 5.2, where concat remains competitive for this reason.

3.3 Baseline: Gate

A single shared gating network reads all the modality embeddings and produces one scalar weight per modality. The weights are normalized, applied to the embeddings, and the weighted sum is passed to the classifier. Gate is the minimal adaptive mechanism: it can learn that one modality is generally more informative than another and can vary that judgment from sample to sample.

However, it cannot vary from label to label because there is a single gate vector for the entire output layer. The network is forced to answer one question, namely which modality matters for this patient, when the task actually asks a different question: which modality matters for this label, for this patient. This per-label flexibility is exactly what the proposed mechanism (Section 3.1) adds, and it is the main structural axis on which the two differ.

3.4 Baseline: Attention (confidence-gated cross-attention)

Unlike the gate and LAG, this mechanism requires one modality to be nominated as the anchor, and the anchor differs by dataset (CXR for Symple-MIMIC and HAIM; chart events for eICU). The anchor embedding forms the query, and the two non-anchor modalities form the key and value.

A confidence gate scales each non-anchor contribution before the attention output is added to the anchor. The mechanism is expressive, but the anchor choice is a dataset-specific hyperparameter that must be set before training. If the anchor is incorrect for a given label, no attention weight can repair it because the label prediction is still built on the anchor representation.

3.5 Baseline: Transformer (confidence-gated transformer encoder)

Structurally close to attention, with the same anchor nomination, but the fused modality set is treated as a token sequence and passed through a full transformer encoder block (multi-head self-attention, feed-forward network, residual connections, and layer normalization). The intent is to allow modalities to interact more richly than a single cross-attention pass allows.

In practice, this is the largest fusion module in the comparison. It has the most parameters of any mechanism tested apart from hyperfusion, and its behavior (Section 5) is the clearest illustration in this paper that more capacity is not a free upgrade when the training set is small and the target is multi-label.

3.6 Baseline: Hyperfusion (hypernetwork-generated classifier)

The same anchor nomination and cross-modal confidence gate as in Sections 3.4 and 3.5 produce a fused context vector $z = [e_{\text{anchor}}; g_B \cdot e_B; g_C \cdot e_C]$, which an MLP projects to a compact conditioning code. A hypernetwork then maps that code to the weights and biases of the final classifier layer, such that the classifier itself is generated per sample rather than learned once.

This is the most parameter-heavy mechanism in this study. It is included because hypernetworks [16] are a recurring proposal for input-conditioned fusion and because it tests the opposite end of the design space from LAG: conditioning the classifier on the sample rather than conditioning the modality mixture on the label.

3.7 Shared training configuration

All models: embed_dim = 512, hidden_dim = 1024, dropout = 0.3, AdamW, linear warmup into ReduceLROnPlateau, gradient clipping, masked weighted BCE loss, minimum-delta early stopping on validation macro-AUROC (to prevent small-validation-set noise from triggering premature stopping), fixed seed with deterministic=True. Identical hyperparameters (aside from dataset-specific batch_size/max_pos_weight, set once from the measured class imbalance) were used for all six mechanisms within each dataset.

4. Experimental setup and reproducibility

4.1 Datasets and multi-label targets

All three datasets are publicly available secondary data hosted by the PhysioNet repository: Symile-MIMIC v1.0.0 (<https://physionet.org/content/symile-mimic/1.0.0/>) [4], HAIM multimodal dataset v1.0.1 (<https://physionet.org/content/haim-multimodal/1.0.1/>) [5], and eICU Collaborative Research Database

v2.0 (<https://physionet.org/content/eicu-crd/2.0/>) [6]. Access to the database requires PhysioNet credentialing. Full access details are provided in the Data Availability Statement. Table 1 summarizes the modality composition, scale, and multi-label targets of each of the three datasets used in this study.

Table 1
Overview of the three datasets and their multi-label targets.

Modalities	Symile-MIMIC	HAIM	eICU-CRD
	ECG (raw waveform, 1D-CNN encoder), CXR (raw image, DenseNet121 encoder), Labs (tabular, MLP encoder)	CXR image embedding (1024-d), radiology/clinical report text embedding (2304-d), chart + lab event embedding (1024-d)	chart-event embedding, lab-event embedding, clinical-note embedding
Modality representation	raw, encoded end-to-end	pre-extracted, frozen	pre-extracted, frozen
Train / test size	10,000 / 750 patients	~ 1,034 / 222 patients	~ 97,500 / ~20,900 patients
Multi-label target	6 CheXpert findings: No Finding, Cardiomegaly, Lung Opacity, Edema, Atelectasis, Pleural Effusion	14 CheXpert findings	8 diagnosis-category flags: cardiovascular, pulmonary, neurologic, renal, gastrointestinal, endocrine, infectious, hematology
Label correlation structure	imaging findings, co-occur on the same CXR	imaging findings, co-occur on the same CXR	organ-system diagnoses, co-occur in the same ICU stay

4.2 Training and evaluation protocol

For the two pre-extracted embedding datasets (HAIM and eICU), each modality was passed through a matched 2-layer MLP adapter (Linear → LayerNorm → ReLU → Dropout → Linear → LayerNorm, embed_dim = 512). The adapter size was selected to match the established encoder capacity for each dataset rather than over- or under-provisioning either. Patient-level stratified splitting was used throughout, not row-level, to prevent data leakage. The eICU, in particular, needs a patient_id split key that is distinct from its per-stay row identifier because some patients have several ICU stays with different diagnosis category outcomes.

Positive-class imbalance is handled with a masked, per-label weighted BCE loss, with pos_weight capped from the measured per-label imbalance of each dataset. Per-label decision thresholds were chosen by an F1-optimal grid search on the validation split only, never on the test split.

Every multi-label result reported below is obtained from four seeds (42, 43, 44, and 45) per mechanism per dataset. Each seed re-seeds model initialization, data shuffling, dropout, and, for HAIM and eICU, the patient-level stratified split. Everything else was fixed: encoder capacity, optimizer, schedule, loss

weighting, thresholding, and metrics. The reported numbers are means over the four runs, with standard deviations, and per-seed values are maintained so that paired differences can be inspected directly.

Under a fixed seed, the results were exactly reproducible on our own hardware; however, we do not claim bit-identical reproducibility in general: cuDNN kernel selection and non-deterministic GPU reductions can change the low-order digits on other setups. The claims in this paper are built to survive this. We report four-seed means with standard deviations, and where two mechanisms are close, we report the per-seed paired difference instead of the aggregate. The 54 additional training runs behind these numbers were executed under one script and one configuration file; therefore, the comparison itself does not depend on run-to-run bit equality.

5. Experimental results: LAG versus baseline fusion mechanisms

This section reports the four-seed experiments described above. Section 5.1 provides the per-dataset means, Section 5.2 aggregates them into a cross-dataset ranking, and Section 5.3 inspects what LAG's per-label attention actually learned.

5.1 Four-seed means by dataset

Each mechanism was trained under four seeds (42/43/44/45; Section 4.2) on each of the three datasets, and all statistics were computed from those runs.

Table 2

Four-seed mean $\pm$ standard deviation, macro-AUROC, by dataset and mechanism. The final row repeats the best single-modality baseline from Table 1

LAG	Symile-MIMIC	HAIM	eICU
	0.7339 $\pm$ 0.0049	0.7118 $\pm$ 0.0163	0.8799 $\pm$ 0.0020
Gate	0.7337 $\pm$ 0.0049	0.6998 $\pm$ 0.0212	0.8815 $\pm$ 0.0019
Hyperfusion	0.7319 $\pm$ 0.0019	0.6873 $\pm$ 0.0338	0.8758 $\pm$ 0.0031
Attention	0.7299 $\pm$ 0.0072	0.7038 $\pm$ 0.0205	0.8760 $\pm$ 0.0026
Concat	0.7293 $\pm$ 0.0127	0.7103 $\pm$ 0.0156	0.8790 $\pm$ 0.0018
Transformer	0.7153 $\pm$ 0.0037 (worst)	0.6493 $\pm$ 0.0507 (worst)	0.8668 $\pm$ 0.0017 (worst)
best unimodal	cxr 0.7334	nlp 0.6678	nlp 0.8600

Three aspects immediately stand out. First, in HAIM, the standard deviation of every mechanism (0.005 to 0.05) is as large as or larger than the gaps between mechanisms. The five non-transformer means span only 0.6873 to 0.7118, a range of 0.023 that a single run on $\sim 1,034$ patients could not resolve. Only the transformer's mean sits clearly below the pack, so HAIM separates the transformer from the rest but does not reliably order the other five.

Second, the two larger datasets exhibited different behaviors. On Symile-MIMIC and eICU, the standard deviations fell to 0.002 to 0.013, and LAG and gate were separated from the rest by margins that held at every seed. Third, the unimodal reference points in Table 2 bound the actual worth of fusion for each dataset. On HAIM, the best single modality reached 0.6678 against LAG's 0.7118, a gain of 0.0440.

On eICU, it reached 0.8600 against gate's 0.8815, a gain of 0.0215. Both gains are an order of magnitude larger than the corresponding seed standard deviations. On Symile-MIMIC, the best single modality (CXR) already reached 0.7334 against 0.7339 for LAG. This difference of 0.0005 sits well inside LAG's own seed standard deviation of 0.0049, and transformer, at 0.7153, falls below the unimodal baseline outright. Symile-MIMIC therefore orders the six mechanisms without showing that any of them earns multimodal complexity. The ranking is a comparison among fusion mechanisms, not evidence that fusion helps.

5.2 Cross-dataset ranking

Aggregating the four-seed means across datasets by average rank and average gap to each dataset's best mechanism:

Table 3
Cross-dataset ranking summary, four-seed means.

Model	Avg. rank (1 = best)	Avg. gap to that dataset's best
LAG	1.33	-0.0005
Gate	2.33	-0.0041
Concat	3.33	-0.0029
Attention	3.67	-0.0058
Hyperfusion	4.33	-0.0107
Transformer	6.00	-0.0319

The proposed mechanism, LAG, has the best average rank and the smallest average gap to the best of the six mechanisms compared. It wins or ties for first on two of three datasets (Symile-MIMIC and HAIM) and places a close second on eICU. The paired per-seed differences against the gate on eICU are 0.0013, 0.0023, 0.0022, and 0.0006; thus, the gate leads by a small margin that is nevertheless consistent at every seed.

The eICU is also the only dataset large enough, at ~ 97,500 patients, for an average difference of 0.0016 AUROC to be a real signal rather than noise. transformer is the only mechanism whose rank is entirely insensitive to seed choice: 6.00, last on all three datasets, and last at every one of the four seeds on Symile-MIMIC and eICU.

5.3 Does LAG's attention differentiate by label?

Averaging LAG's learned attention matrix $A[\text{label}, \text{modality}]$ over each dataset's full test set:

eICU (eight labels): Real and clinically sensible differentiation.

The labels `dx_renal` and `dx_hematology` are the two diagnostic categories most directly established by laboratory values: creatinine and BUN for renal function and CBC and coagulation panels for hematologic disease. They are also the only two labels where LAG shifts attention toward the lab modality, relative to every other label. This routing was not supervised. This emerges purely from the multi-label training objective.

Table 4
LAG's learned attention weights by label, eICU (eight labels).

Label	chart	lab	nlp	Argmax
dx_cardiovascular	0.293	0.271	0.436	nlp
dx_pulmonary	0.296	0.311	0.392	nlp
dx_neurologic	0.277	0.289	0.434	nlp
dx_renal	0.272	0.367	0.361	lab
dx_gastrointestinal	0.281	0.313	0.406	nlp
dx_endocrine	0.271	0.349	0.380	nlp
dx_infectious	0.284	0.314	0.403	nlp
dx_hematology	0.282	0.361	0.357	lab

Symile-MIMIC (six labels): real but modest differentiation, correctly CXR-anchored.

All six labels correctly attended to CXR most, and the degree scales sensibly with how radiographically distinctive each finding is. Edema and Pleural Effusion are large and visually unambiguous, and they pull CXR attention to 0.38 and 0.39, respectively. No Finding and Lung Opacity are subtler, and they remain near 0.34.

Table 5
LAG's learned attention weights by label, Symile-MIMIC
(six labels).

Label	ecg	cxr	labs	Argmax
No Finding	0.326	0.341	0.334	cxr
Cardiomegaly	0.325	0.348	0.326	cxr
Lung Opacity	0.328	0.341	0.331	cxr
Edema	0.304	0.382	0.314	cxr
Atelectasis	0.312	0.370	0.318	cxr
Pleural Effusion	0.301	0.390	0.309	cxr

HAIM (14 labels): a null result, where the attention collapses to uniform.

Across all 14 labels, LAG's attention remained within 0.327 to 0.344 for every modality, with a standard deviation of approximately 0.002 to 0.005 per modality across labels. This is functionally indistinguishable from an unweighted average. With only ~ 1,034 training patients spread across 14 labels, there is apparently not enough per-label signal for the 14 query vectors to learn anything beyond the shared label-agnostic average. On this particular seed, raw accuracy placed concat above LAG on HAIM (Section 5.1).

The four-seed mean (Section 5.2) places LAG fractionally above concat in this dataset. The finding should therefore be read as "LAG's interpretability did not pay off in this run," not as "LAG lost to a simpler, unweighted-average mechanism," because the second claim did not survive the correction in Section 5.2. However, interpretability collapse remains a concern for this run. Interpretability, like raw accuracy, is evidently data-hungry, and does not come for free from the architecture. Whether it collapses on every seed for a dataset of this size remains untested (Section 7).

6. Discussion and recommendations

Before choosing a fusion mechanism, check that fusion beats the strongest single modality on your own dataset. This is a different question from ranking, and our three datasets do not answer it in the same way. Fusion cleared the best unimodal baseline by 0.0440 macro-AUROC on HAIM and 0.0215 on eICU, but only by 0.0005 on Symile-MIMIC. In that dataset, the margin was smaller than LAG's own seed standard deviation, and the transformer performed worse than the CXR encoder used alone (Section 5.1). A mechanism ranking is still meaningful there, but it ranks options, none of which has yet justified its extra complexity. The reference point for the decision to fuse is the best unimodal model, not the best fusion mechanism. It is also cheap to compute, so there is no reason to skip it.

For multi-label clinical diagnoses, we recommend the proposed mechanism as the default. Under the four-seed means, LAG has the best average rank (1.33) and the smallest average gap to best (– 0.0005)

among the six mechanisms (Section 5.2). It carries no anchor-modality hyperparameter to re-derive, its added cost is $L \times d$ parameters (Section 3.1), and it is the only mechanism in the comparison that also offers per-label attribution when there is sufficient data to support it (Section 5.3).

The honest boundary of this recommendation is the eICU. Gate leads there at every one of the four seeds, on the one dataset large enough for a 0.0016 AUROC margin to be trustworthy (Section 5.2). Therefore, we recommend LAG as the default and flag gate as a legitimate alternative worth measuring on any new dataset. The two are close enough that dataset-specific measurements, not a fixed ranking decided in advance, should make the final call. Where they are statistically tied, the per-label attribution LAG provides a second reason to prefer it.

Avoid transformers as a multi-label default. This is the only recommendation in this paper that requires no caveat. Under multi-label evaluation, it is the worst mechanism on every dataset tested, and on the two datasets with variance tight enough to resolve it, at every one of the four seed runs (Section 5.2). This is the most robust finding of this study.

Hyperfusion was treated as unresolved rather than ruled out and checked per dataset. hyperfusion remains consistently mid-to-weak across all three datasets and all four seeds. This study does not disqualify the use of HAIM-sized datasets, but it cannot be recommended or avoided based on a single evaluation.

7. Limitations

The findings above come from one specific evaluation design, and that design bounds how far they generalize to other designs. We list below the constraints a reader should consider before applying our recommendations, ordered from the evaluation protocol to the datasets to the mechanism itself.

- Four seeds bound the uncertainty only coarsely. We report means, standard deviations, and raw per-seed paired differences (Sections 5.1 and 5.2) rather than a significance test because with $n = 4$, a t-test would invite the false precision that this study argues against. The gate and LAG comparison on eICU (Section 5.2) is the only one whose direction holds across all four seeds on a dataset large enough ($\sim 97,500$ patients) to trust. On Symile-MIMIC fusion, the best single modality is cleared by only 0.0005 macro-AUROC (Section 5.1); thus, the conclusion rests on HAIM and eICU. All other close comparisons should be read as unranked.
- The datasets differ in more ways than we can separate them. HAIM and eICU use pre-extracted, frozen embeddings, while Symile-MIMIC trains its ECG and CXR encoders end to end; therefore, results may differ for architectures trained on raw inputs throughout. The targets also differ in terms of label count, correlation structure, and dataset identity. Therefore, we cannot cleanly separate “more labels hurt a mechanism” from “this specific dataset hurts a mechanism” for any per-dataset pattern reported here, including HAIM’s collapse of LAG interpretability (Section 5.3).

- LAG’s attention interpretability of LAG (Section 5.3) was evaluated qualitatively on the seed-42 models only, not re-run across the additional three seeds, and not checked against a ground-truth modality-relevance annotation (none exists for any of these three datasets). Whether HAIM’s specific attention-collapse result would look different for other seeds is untested and, given HAIM’s per-seed spread (Section 5.1), it should not be assumed to be stable.

These limitations do not change the mechanism ranking reported here, but they define its scope. The results suggest a default choice for multi-label clinical fusion on datasets similar to the three tested here, rather than a fixed order that is automatically applied to a new dataset.

8. Conclusion

This study proposes label-adaptive gating (LAG), a multimodal fusion mechanism that resolves modality importance per label rather than per sample or dataset. It cross-attends one learned query vector per label over the stacked modality embeddings, requires no anchor-modality hyperparameter, and adds $L \times d$ parameters (Section 3.1).

We tested the proposal against five representative fusion baselines and per-modality unimodal models, holding the encoder capacity, training protocol, and evaluation fixed, on three independent and genuinely multi-label clinical datasets. Under four-seed means, LAG obtained the best average rank (1.33 of 6) and the smallest average gap to each dataset's best mechanism (-0.0005 macro-AUROC) across the Symile-MIMIC, HAIM, and eICU datasets. It is also the only mechanism that offers per-label modality attribution. This attribution is genuine and clinically sensible when a label has sufficient training data, and we report it honestly as an uninformative collapse when it does not (Section 5.3). Where the proposal does not win, we say so: scalar gating leads on eICU at every seed, by a small margin, on the one dataset that is large enough to resolve it (Section 5.2).

Among the baselines, one finding is unambiguous and low-variance: the confidence-gated transformer encoder is the worst multi-label mechanism tested, last on all three datasets, and last at every seed on the two large enough for a per-seed ordering.

One direction follows directly from these results: LAG's per-label queries are static, whereas the scalar gate we compared against is conditioned on the sample. The eICU results show that the two conditioning axes are not interchangeable; therefore, a mechanism that conditions both the label and the sample is the natural next step. We excluded this design so that any observed effect could be attributed to label granularity alone.

Declarations

Clinical trial number

Not applicable. This study was a retrospective secondary analysis of existing de-identified datasets and was not a clinical trial.

Ethics approval:

Not applicable. No human participants were involved in this study, and it did not include surveys, questionnaires, or feedback processes. This study was conducted using publicly available, de-identified datasets (Symile-MIMIC, HAIM, and eICU Collaborative Research Database). All three databases contained anonymized patient data and were collected with institutional review board approval from the respective institutions. As this study involved a secondary analysis of fully de-identified data, additional ethical approval and informed consent were not required.

Consent to Participate in the study

Not applicable. All data were drawn from existing de-identified databases, and consent was obtained from the original data providers.

Consent to Publish:

Not applicable.

Funding declaration

This research was funded by Tra Vinh University (TVU) under Grant Contract 50/2026/HĐ.HĐKH&ĐT–ĐHTV.

Author Contribution

Here is the formal Author Contributions Statement for your manuscript, structured following standard CRediT (Contributor Roles Taxonomy) guidelines required by Discover Artificial Intelligence and Springer Nature journals. Author Contributions Statement Van-Thien Tran: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Project administration. Thanh-Nghi Doan: review & editing, Supervision. Song-Toan Tran: review & editing, Supervision, Corresponding Note: All authors have read and agreed to the published version of the manuscript.

Data availability statement:

The datasets analysed during the current study are available in the PhysioNet repository: Symile-MIMIC v1.0.0, <https://physionet.org/content/symile-mimic/1.0.0/> [4]; HAIM multimodal dataset v1.0.1, <https://physionet.org/content/haim-multimodal/1.0.1/> [5]; and eICU Collaborative Research Database

v2.0, <https://physionet.org/content/eicu-crd/2.0/> [6]. Access requires PhysioNet credentialing and completion of the required human subjects research training course. No new data were generated.

References

1. Sleeman IV, Kapoor WC, Ghosh R. P.: Multimodal classification: Current landscape, taxonomy and future directions. *ACM Computing Surveys*, 55(7), Article 150 (2022).
<https://doi.org/10.1145/3543848>
2. Al-Zoghby AM, Ebada AI, Saleh AS, Abdelhay M, Awad WA. A comprehensive review of multimodal deep learning for enhanced medical diagnostics. *Computers Mater Continua*. 2025;84(3):4155–93.
<https://doi.org/10.32604/cmc.2025.065571>.
3. Hayat N, Geras KJ, Shamout FE. MedFuse: Multi-modal fusion with clinical time-series data and chest X-ray images. In: *Proceedings of Machine Learning Research (ML4H)* (2022).
4. Saporta A, Puli AM, Goldstein M, Ranganath R. Symile-MIMIC: a multimodal clinical dataset of chest X-rays, electrocardiograms, and blood labs from MIMIC-IV (version 1.0.0). *PhysioNet* (2025).
<https://doi.org/10.13026/3vvj-s428>
5. Soenksen LR, Ma Y, Zeng C, Boussieux LDJ, Villalobos Carballo K, Na L, Wiberg H, Li M, Fuentes I, Bertsimas D. Code for generating the HAIM multimodal dataset of MIMIC-IV clinical data and x-rays (version 1.0.1). *PhysioNet* (2022). <https://doi.org/10.13026/3f8d-qe93>
6. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. eICU Collaborative Research Database (version 2.0). *PhysioNet* (2019). <https://doi.org/10.13026/C2WM1R>
7. Zhu H, Cui Q, Li J, Zhang L. A multi-scale expert gating network with cross-modal attention for multimodal medical image fusion. *Biomedical Signal Processing and Control*; 2026.
8. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. *Adv Neural Inf Process Syst* (2017).
9. Deepsahith K, Shashank BN, Kumar S, Alphonse S, Subburaj T, Subramanian S. Graph-enhanced multimodal fusion of vascular biomarkers and deep features for diabetic retinopathy detection. *Front Artif Intell*. 2026;8:1731633. <https://doi.org/10.3389/frai.2025.1731633>.
10. Zhou HY, Yu Y, Wang C, Zhang S, Gao Y, Pan J, Shao J, Lu G, Zhang K, Li W. A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. *Nat Biomedical Eng*. 2023;7:743–55. <https://doi.org/10.1038/s41551-023-01045-x>.
11. Khader F, Kather JN, Mueller-Franzes G, Wang T, Han T, Arasteh T, Hamesch S, Bressemer K, Haarbuerger K, Stegmaier C, Kuhl J, Nebelung C, Truhn S. Medical transformer for multimodal survival prediction in intensive care: integration of imaging and non-imaging data. *Sci Rep*. 2023;13:10666. <https://doi.org/10.1038/s41598-023-37835-1>.
12. Perez E, Strub F, De Vries H, Dumoulin V, Courville A. FiLM: Visual reasoning with a general conditioning layer. In: *AAAI* (2018). <https://doi.org/10.1609/aaai.v32i1.11671>

13. Zadeh A, Chen M, Poria S, Cambria E, Morency LP. Tensor fusion network for multimodal sentiment analysis. arXiv preprint arXiv. 2017. <https://doi.org/10.48550/arXiv.1707.07250>. :1707.07250.
14. Liu Z, Shen Y, Lakshminarasimhan VB, Liang PP, Zadeh A, Morency LP. Efficient low-rank multimodal fusion with modality-specific factors. arXiv preprint arXiv. 2018. <https://doi.org/10.48550/arXiv.1806.00064>. :1806.00064.
15. Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE. Adaptive mixtures of local experts. *Neural Comput.* 1991;3(1):79–87. <https://doi.org/10.1162/neco.1991.3.1.79>.
16. Ha D, Dai A, Le QV. Hypernetworks. arXiv preprint arXiv:1609.09106 (2017). <https://doi.org/10.48550/arXiv.1609.09106>
17. Johnson AE, Pollard TJ, Greenbaum NR, Lungren MP, Deng CY, Peng Y, Lu Z, Mark RG, Berkowitz SJ, Horng S. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019). <https://doi.org/10.48550/arXiv.1901.07042>
18. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, Marklund H, Haghighi B, Ball R, Shpanskaya K, et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *AAAI.* 2019;33(01):590–7. <https://doi.org/10.1609/aaai.v33i01.3301590>.
19. Johnson AE, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, Pollard TJ, Moody B, Gow B, Lehman LWH, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data.* 2023;10:1. <https://doi.org/10.1038/s41597-022-01899-x>.
20. Soenksen LR, Ma Y, Zeng C, Boussieux L, Villalobos Carballo K, Na L, Wiberg HM, Li ML, Fuentes I, Bertsimas D. Integrated multimodal artificial intelligence framework for healthcare applications. *npj Digit Med.* 2022;5:149. <https://doi.org/10.1038/s41746-022-00689-4>.
21. Pollard TJ, Johnson AE, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Sci Data.* 2018;5:180178. <https://doi.org/10.1038/sdata.2018.178>.
22. Zhang T, Han L, Gao Y, Wang X, Beets-Tan R, Mann R. Predicting molecular subtypes of breast cancer using multimodal deep learning and incorporation of the attention mechanism. In: *Medical Imaging with Deep Learning (MIDL). Short Paper* (2021).
23. Vaezi Joze HR, Shaban A, Iuzzolino ML, Koishida K. MMTM: Multimodal transfer module for CNN fusion. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13289–13299 (2020). <https://doi.org/10.1109/CVPR42600.2020.01330>
24. Pölsterl S, Wolf TN, Wachinger C. Combining 3D image and tabular data via the dynamic affine feature map transform. Volume 12905. In: *MICCAI; 2021*. pp. 688–98. https://doi.org/10.1007/978-3-030-87240-3_66.
25. Kanthavel R, Dhaya R, Jayanthiladevi A, Rutakemwa MM. Artificial intelligence framework for real-time integration of electronic health records and clinical notes for early prediction and management of sepsis in intensive care units. *Discover Artif Intell.* 2026;6(1):801.
26. Haripriya T, Dharmarajan K, Chowdhury S, Panigrahi A, Pati A, Sahu B. A hybrid hierarchical multimodal learning with autoencoders, vision transformers and deep belief networks for liver

- cancer detection. *Discover Artif Intell.* 2026;6(1):824.
27. Jain E, Kukreja V, Kaushik P, Jain AK, Bali MS. A reliability-aware multi-feature fusion framework using fuzzy logic for interpretable ear disease classification from otoscopic images. *Discover Artif Intell.* 2026;6(1):828.
28. Zhao F, Zhang C, Geng B. Deep multimodal data fusion. *ACM Computing Surveys*, 56(9), Article 216 (2024). <https://doi.org/10.1145/3649447>
29. Xiao L, Huang X, Chen B, Jing L. Label-specific document representation for multi-label text classification. *Association for Computational Linguistics, Hong Kong*; 2019. pp. 466–75. <https://doi.org/10.18653/v1/D19-1044>.
30. Vu T, Nguyen DQ, Nguyen A. A label attention model for ICD coding from clinical text. In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)*, pp. 3335–3341 (2020).

Figures

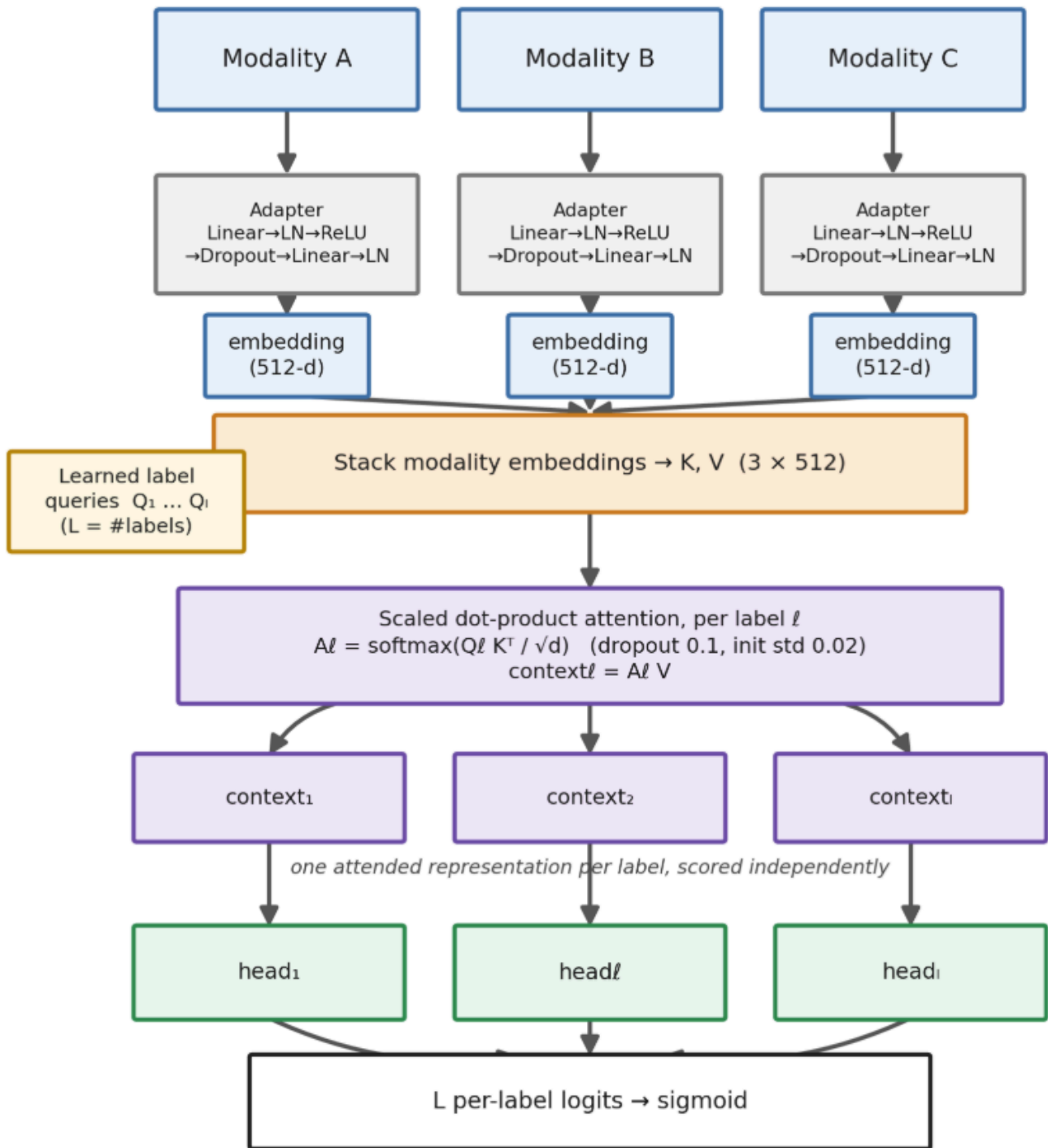


Figure 1

LAG: One learned query vector per label cross-attends over the stacked modality embeddings, producing a label-specific attended representation.

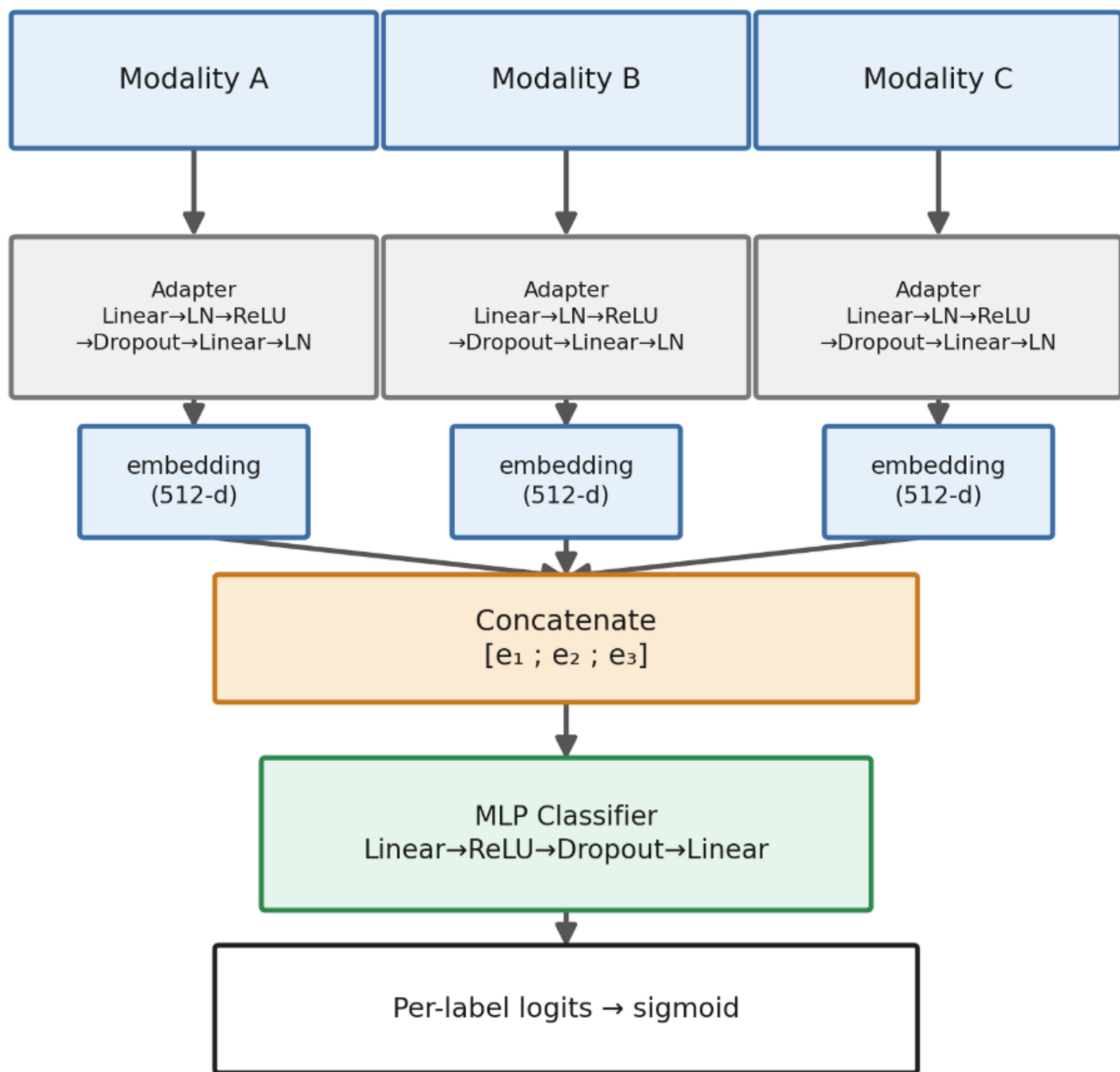


Figure 2

Concat: Modality embeddings are concatenated and passed directly to an MLP classifier with no learned inter-modality weighting.

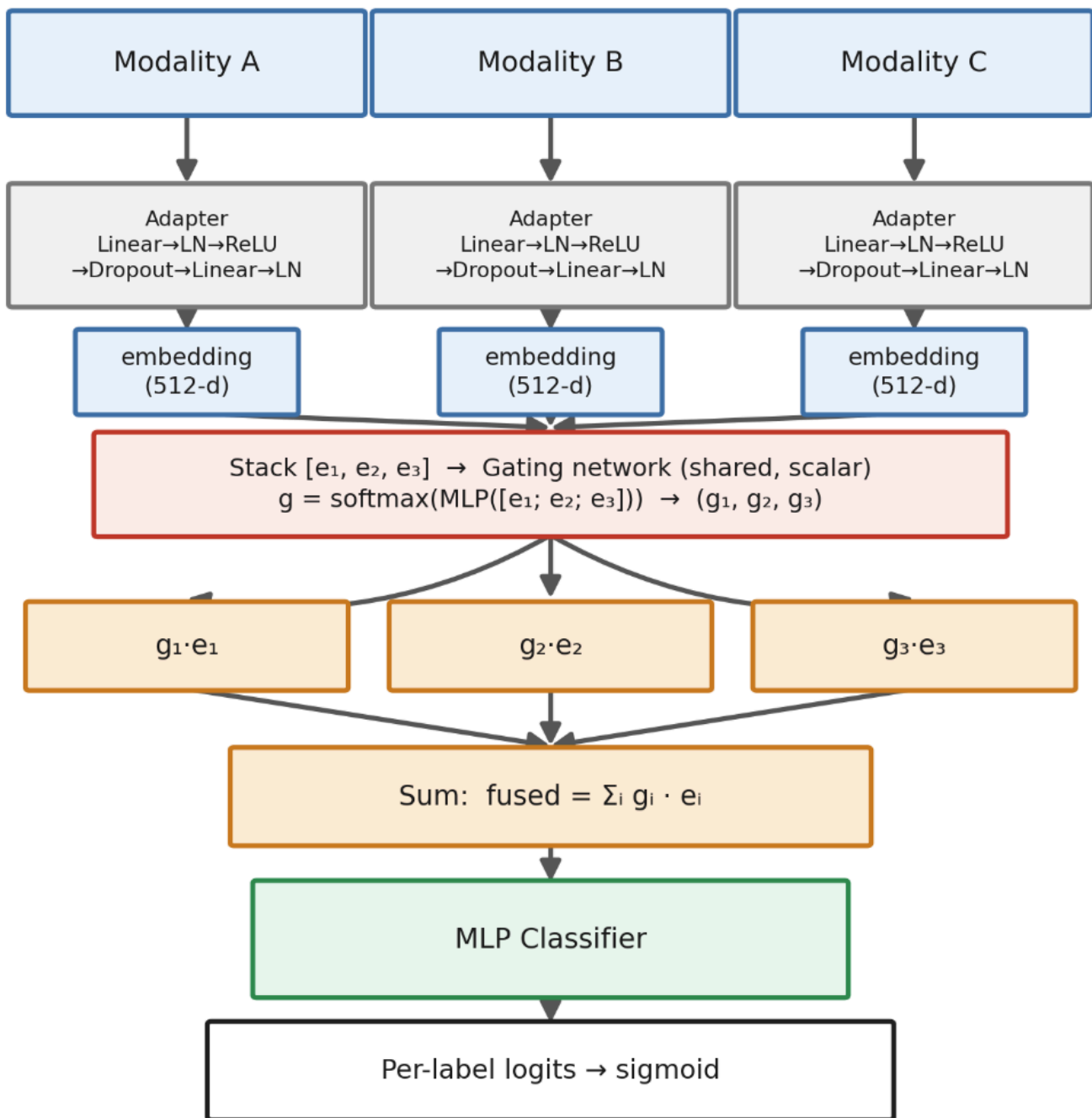


Figure 3

Gate: A single shared gating network produces one scalar weight per modality, which is applied identically across every label.

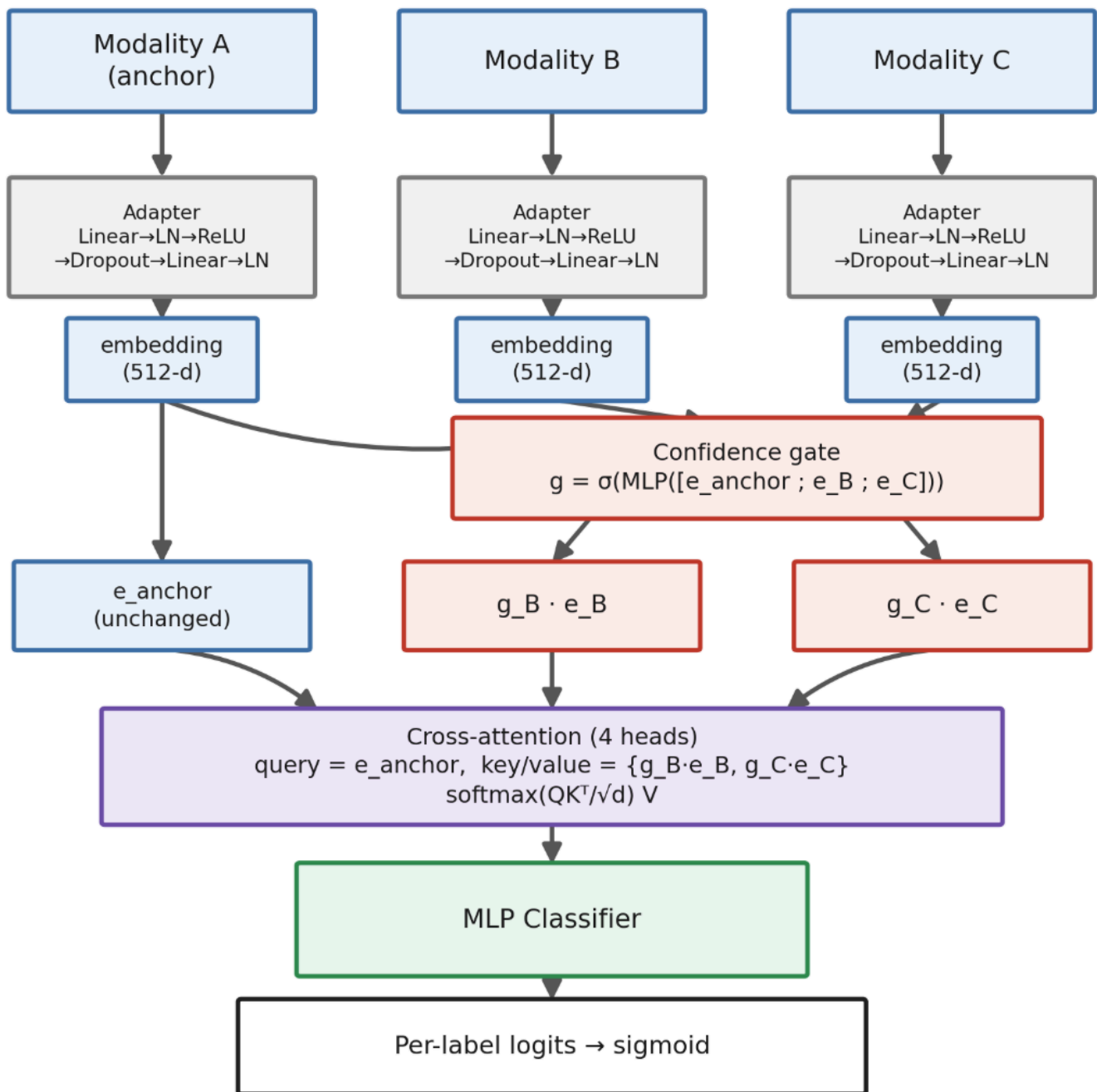


Figure 4

Attention: Confidence-gated cross-attention with one modality nominated in advance as the anchor query.

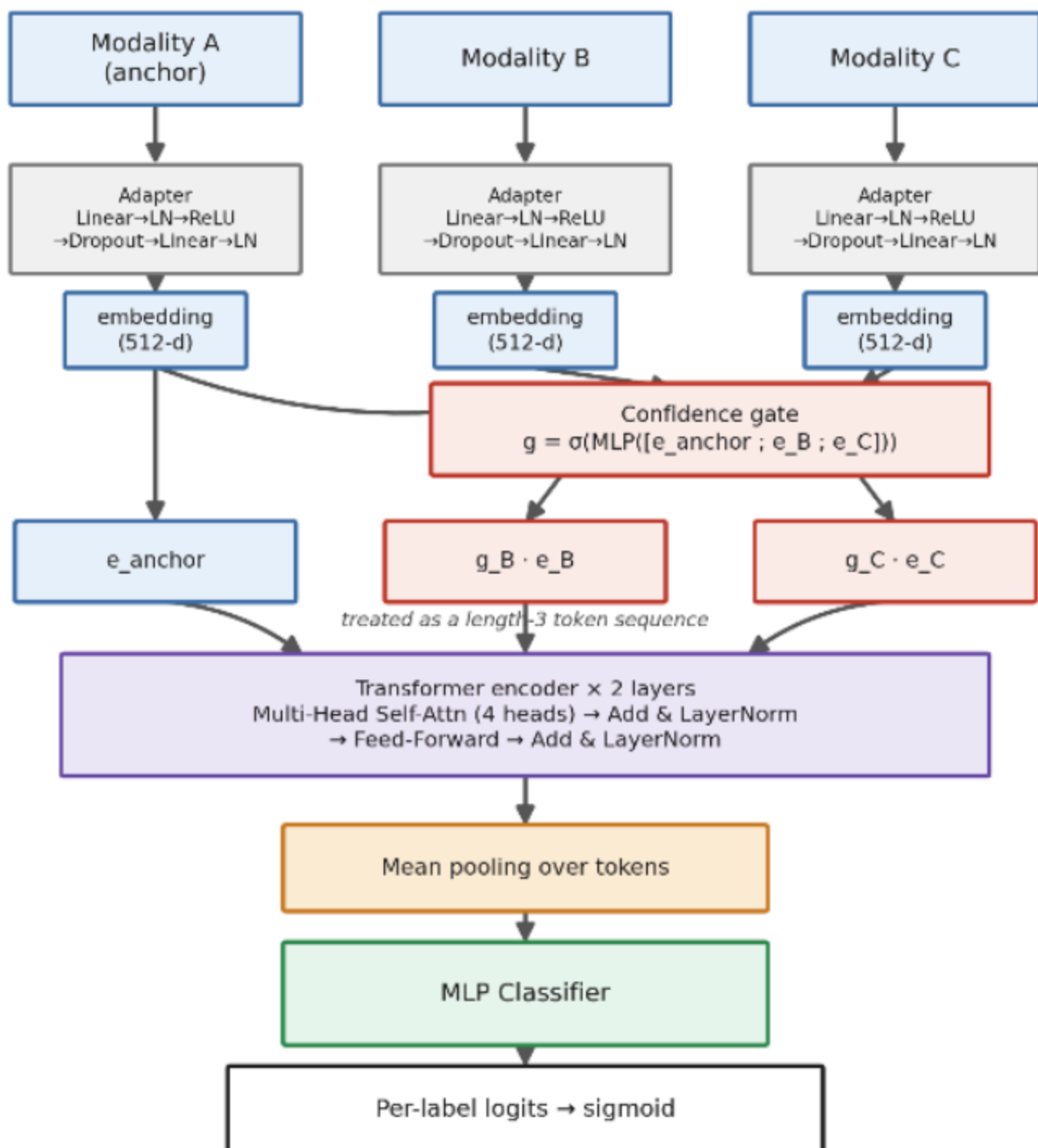


Figure 5

Transformer: The gated modality set is treated as a token sequence and passed through a 2-layer transformer encoder.

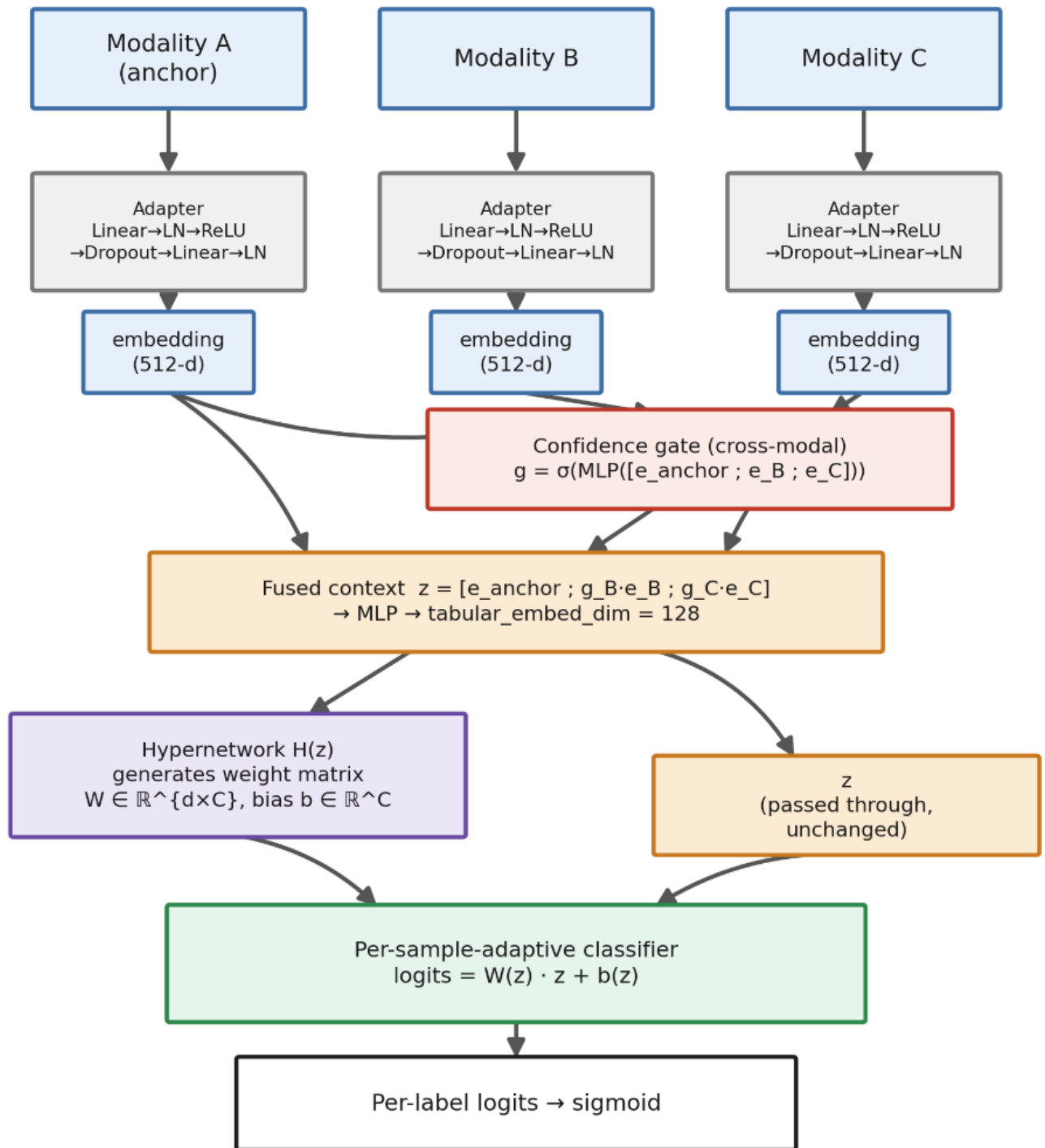


Figure 6

Hyperfusion: A hypernetwork generates the final classifier weights conditioned on the fused representation, thereby producing a per-sample adaptive decision boundary.