WHO GETS TO CONSTRAIN THE DECISION? AUTHORITY-CONDITIONED GRAPH FUSION FOR CLINICAL MULTIMODAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal feature fusion combines heterogeneous evidence for structured decision-making, but clinically motivated scenarios raise a fundamental question: when predictive strength, reliability, and evidential authority conflict, which evidence source should constrain the final output? We formulate this problem as an authority-asymmetric multimodal learning process and propose the Authority-Consistent Fusion (ACF) framework that treats authority as a task-dependent specification. ACF includes three components: a label-free Authority Specification Compiler to derive explicit authority constraints, an Authority-Protected Graph Mixer to prevent unauthorized information propagation, and an Authority-Consistent Projection to enforce admissible outputs under missing, invalid, and conflicting evidence. Under explicit assumptions, we establish relation-preservation, fixed-parameter non-interference, and conditional squared-loss guarantees for the proposed framework. Experiments on the UTSW MRI-to-semantic retrieval benchmark demonstrate that ACF achieves comparable calibration to information-matched learned joint graph fusion, with a clean Brier score difference within 7.0×10^{-5}. Controlled evaluations further show that ACF maintains a zero S1 authority-violation rate under activated relations and incurs Brier increases of only 0.084–0.189 under incorrect authority specifications. Additional matched experiments attribute 6.49 and 2.48 mAP improvements to the evaluated geometric and communication configurations, respectively. ACF establishes evidential authority as an explicit decision variable and provides a principled foundation for reliable and accountable multimodal decision-making when heterogeneous evidence sources disagree.

1 INTRODUCTION

Multimodal feature fusion combines heterogeneous evidence to support structured decisions, yet the available sources may differ along conceptually distinct axes. Prior multimodal work separates how much a source contributes to task prediction from whether its current observation is trustworthy under uncertainty, which we call **predictive utility** and **reliability**, respectively (Han et al., 2021; Zhang et al., 2023; Gao et al., 2024). A modality may therefore be predictive on the training distribution yet unreliable or unavailable for a particular instance (Wang et al., 2023). Neither quantity, however, determines whether the source is entitled to constrain a particular decision relation. Drawing on the domain-specific notion of epistemic authority and the clinical distinction between evidence certainty and recommendation strength (Zagzebski, 2012; GRADE Working Group, 2004; Guyatt et al., 2025), we define **evidential authority** as the task-conditioned entitlement of a valid and available source to constrain a specified relation in the output. This is an operational learning variable, not a claim that an authorized source is universally more accurate, reliable, or informative. This raises the central question of this work: **when predictive utility, reliability, and evidential authority disagree, which evidence source gets to constrain the final decision?**

Existing approaches model several forms of multimodal asymmetry, but these asymmetries do not by themselves answer this question. Attention, gating, and modality weighting estimate how strongly each source should contribute to prediction, while recent cross-attention systems allow one modality to condition the representation of another (Zhang et al., 2025). Quality-aware and uncertainty-aware

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.

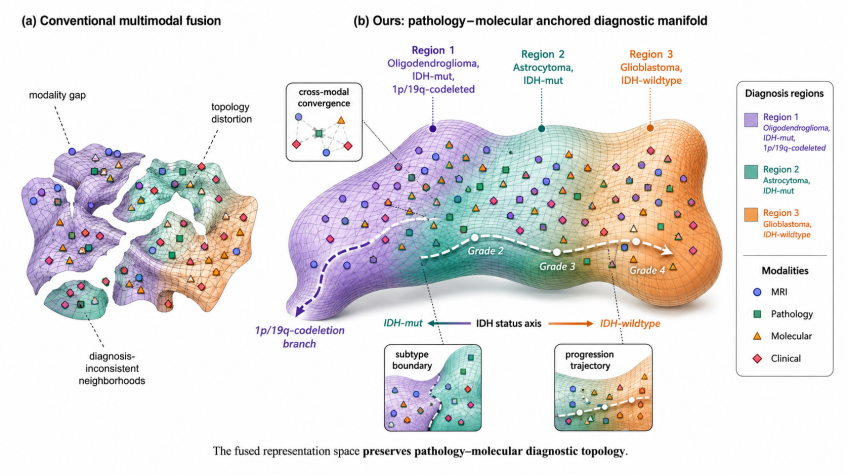


Figure 1: Intended effect of authority-consistent fusion. Conventional fusion may bring modalities together while distorting diagnostic neighborhoods (a). The proposed framework aims to organize complementary evidence around pathology–molecular semantic anchors, preserving subtype boundaries and grade-related structure in the fused representation (b). This is a conceptual illustration, not a measured embedding.

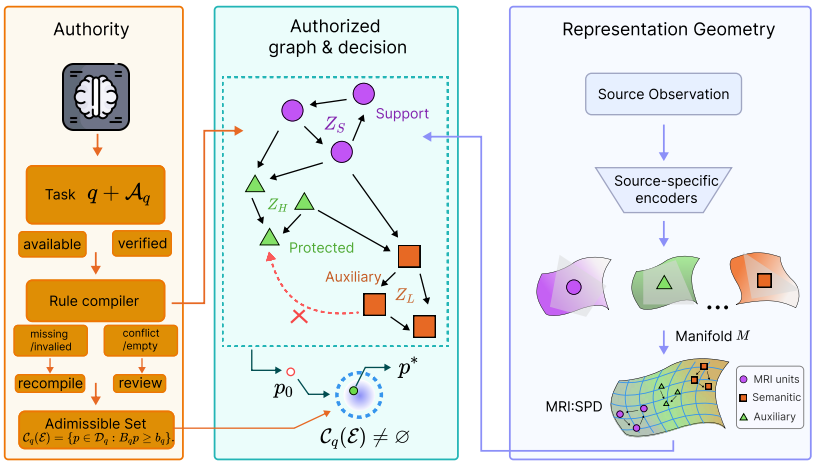


Figure 2: Authority-consistent graph fusion. Source-specific encoders represent the observations, while task-conditioned authority rules define permitted graph communication and the admissible output set. The fused prediction is projected onto this set; missing, invalid, or conflicting evidence follows separate handling paths.

fusion methods model observation quality or confidence (Zhang et al., 2023; Gao et al., 2024), and recent multi-view methods incorporate expert knowledge to improve uncertainty estimation and feature-level interpretation (Liang et al., 2025). Graph-based fusion represents which sources communicate and how their structures interact (Shi et al., 2025). Separately, recent constraint-based learning methods encode requirements on model outputs and can guarantee satisfaction for broad classes of domain constraints (Goyal et al., 2024). These mechanisms can express unequal influence, directional interaction, confidence, expert knowledge, or admissible outputs. Recent clinical systems further explore patient-specific modality selection, entropy-guided image fusion, image–tabular assessment, early-horizon prediction, and resource-aware multimodal screening (Qu et al., 2026; Sagar, 2026; Ochs et al., 2026; Bakumenko et al., 2026; Chan et al., 2026). However, unequal influence does not identify which evidence relation is entitled to constrain a task-specific decision. A large attention weight may reflect predictive utility; low uncertainty may reflect reliability; and a

learned edge may reflect useful communication. None of these quantities, in isolation, establishes that the corresponding source has been authorized to impose a relation on the output.

We therefore formulate **authority-asymmetric multimodal learning**, in which evidential authority is represented as an explicit task variable. Authority is **task-conditioned** because the same evidence may constrain one target but remain auxiliary for another; **relation-specific** because authorization applies to a particular decision relation rather than to an entire modality; **validity- and availability-dependent** because a relation is withdrawn when its supporting evidence is invalid or missing; and **externally specified** because it originates from the task protocol rather than being inferred retrospectively from test performance. This variable cannot generally be replaced by the reliability estimates or learned fusion weights used in recent robust multimodal methods (Zhang et al., 2023; Gao et al., 2024; Liang et al., 2025). A minimal example makes the distinction explicit. Let two independent binary sources X_1 and X_2 have identical marginal reliability, let $A \in \{1, 2\}$ specify which source is authoritative, and define $Y = X_A$. A predictor observing X_1 and X_2, but not A, incurs Bayes classification error $1/4$, whereas observing A permits zero error. Thus, even complete access to equally reliable observations does not reveal which relation should govern the decision. A suitable fusion mechanism must consequently compile authority into admissible information dependencies, prevent unauthorized sources from rewriting protected states, and ensure that the output obeys activated relations while auxiliary evidence remains useful within the residual feasible set.

These requirements lead to **Authority-Consistent Fusion (ACF)**, an authority-conditioned graph fusion framework that constrains both computation and decision. The **Authority Specification Compiler** converts a label-free authority specification, together with validity and availability states, into an admissible communication graph and a set of active output constraints. Output constraints alone are insufficient because an unauthorized source could alter an authorized representation through direct or multi-hop message passing before the constraint is applied. The **Authority-Protected Graph Mixer** therefore restricts the information ancestors of protected states while allowing admissible auxiliary communication elsewhere in the graph. Computational protection alone is also insufficient because an unconstrained prediction head may still violate an activated evidence relation. The **Authority-Consistent Projection** maps the unconstrained prediction onto the corresponding authority-consistent feasible set and defines explicit behavior for missing, invalid, and conflicting evidence. Under stated assumptions, this construction provides activated-relation preservation, fixed-parameter computational non-interference, and a conditional squared-loss guarantee when the specified feasible set is correct. These guarantees do not presume that an authority policy is infallible: an incorrect specification can exclude the appropriate decision and increase risk. This failure mode is part of the formulation and must be measured rather than hidden by aggregate predictive performance. Figure 1 highlights the resulting design objective: complementary modalities should be organized around task-relevant semantic anchors while protected diagnostic neighborhoods and decision relations remain intact, rather than being collapsed by unconstrained fusion.

We evaluate these claims through interventions designed to separate authority from its common proxies. In 81 UTSW MRI-to-semantic retrieval runs, geometry and communication contributed $+6.49$ and $+2.48$ mAP points, respectively, separating these effects from learned edge allocation. Across 510 controlled authority experiments, ACF matched the clean predictive utility of an information-matched learned-joint graph model within 7.0×10^{-5} Brier score and attained zero S1 authority-violation rate under the activated relations. Deliberately incorrect authority policies increased Brier score by 0.084–0.189, empirically exposing the cost of invalid constraints rather than treating authority as a universally beneficial prior.

Contributions.

- **Authority as an explicit task variable.** We formulate authority-asymmetric multimodal learning to answer which evidence source is entitled to constrain a decision. The formulation separates evidential authority from predictive utility and reliability, conditions it on the target relation, and specifies how validity, availability, and conflict determine the active decision constraints.
- **From authority to admissible computation and decisions.** We introduce ACF, which compiles a label-free authority specification into an admissible communication graph, protects authorized states from unlicensed information paths, and projects predictions onto an

authority-consistent feasible set. We establish relation-preservation, non-interference, and conditional risk guarantees, together with the consequences of incorrect constraints.

- **Identifying authority beyond proxy effects.** We develop matched interventions and information-matched baselines that vary authority separately from reliability, modality dominance, communication, geometry, and learned edge allocation. This protocol distinguishes predictive utility from structural compliance and tests both correct and incorrect specifications instead of inferring authority from attention weights, learned topology, or task performance.

2 Related Work

Geometric and manifold-based fusion. Multimodal representation learning maps heterogeneous observations into comparable spaces through manifold alignment, shared latent geometry, and Riemannian modeling of structured statistics before downstream fusion (Ham et al., 2005; Arsigny et al., 2007; Poklukar et al., 2022). Recent work examines when cross-modal alignment emerges and uses geometry-aware objectives to preserve shared structure without erasing modality-specific variation (Tang et al., 2024; Tjandrasuwita et al., 2025). We use manifold gluing in this narrower sense: constructing controlled connections between otherwise distinct representations. Such geometry alone determines which relations can be represented, aligned, or connected, but not which relation is entitled to constrain a task-specific decision. Recent clinical and biological alignment systems similarly organize representations through phenotype-guided, vision–language, or group-level cross-modal correspondence (Nizam et al., 2026; Leem et al., 2026; Ji et al., 2026); these alignment objectives provide semantic structure without themselves specifying authority.

Graph learning for cross-modal communication. Graph-based models have progressed from predefined connectivity to learned, dynamic, directional, and sample-dependent structures that adapt message passing to downstream tasks for heterogeneous relational data (Yoon et al., 2023; Campbell et al., 2023; Manenti et al., 2025). These models jointly optimize communication paths and edge allocation, often treating predictive or stable edges as meaningful structure. Yet point-prediction objectives need not identify latent relations: minimizing a downstream prediction loss does not in general guarantee recovery of the underlying graph or its uncertainty (Manenti et al., 2025). Learned topology therefore primarily specifies useful communication, whereas authority specifies which information dependencies are admissible for a particular decision relation. This distinction also applies to uncertainty-aware patient–modality graphs, which adapt message passing according to estimated aleatoric reliability but do not externally authorize decision relations (Gong et al., 2026).

Multimodal fusion under utility, reliability, and constraints. Multimodal fusion models asymmetry through cross-modal attention, gating, modality weighting, and directional interaction at the representation and prediction levels (Zhang et al., 2025; Gao et al., 2024). Uncertainty-aware and trusted methods adjust contributions using estimated reliability, while recent approaches incorporate expert constraints or isolate modality-specific pathways to mitigate cross-modal interference (Gao et al., 2024; Liang et al., 2025; Li et al., 2026). Recent missing-modality methods combine evidential uncertainty, vacuous evidence, conformal calibration, or probabilistic feature imputation to control fusion under incomplete observations (Xing et al., 2026; Moayedikia, 2026; Shahid et al., 2026). These mechanisms improve reliability estimation and availability handling, but neither calibrated uncertainty nor imputation confidence grants a source authority. Recent constraint-satisfaction methods further restrict admissible model outputs during learning (Goyal et al., 2024), while annotation-assisted multimodal policy learning uses expert information to preserve decision-relevant structure (Arno & Demeester, 2026). These methods can encode unequal influence, trust, communication, supervision, or decision rules. Authority-conditioned fusion additionally requires one specification to jointly govern both admissible computation and admissible decisions.

3 METHOD

3.1 OVERVIEW

Authority-Consistent Fusion (ACF) treats evidential authority as an externally specified, task-conditioned variable rather than a quantity inferred from attention, reliability, or predictive performance. ACF constrains both the dependencies used to construct a prediction and the relations satisfied by the returned output. ASC maps $\mathcal{E}$ to $(\mathcal{G}_q, \mathcal{C}_q, \sigma)$; APGM produces $\boldsymbol{p}_0$; and ACP returns $\boldsymbol{p}^\star$. The Authority Specification Compiler (ASC) activates relation-specific permissions and constraints. The Authority-Protected Graph Mixer (APGM) prevents unauthorized inputs from rewriting protected states, while the Authority-Consistent Projection (ACP) enforces the activated relations at the output. Figure 2 summarizes this compile–mix–project pipeline: ASC converts the authority specification and evidence state into admissible communication and output constraints, APGM performs protected fusion, and ACP returns the nearest feasible prediction or invokes an explicit failure-handling path. This separation prevents a learned weight or edge from silently serving as a proxy for authorization, computation, and decision constraints. The source encoder is replaceable. We instantiate it with mean-augmented SPD moments, but neither the geometry nor learned edge weights define authority.

3.2 AUTHORITY-ASYMMETRIC MULTIMODAL LEARNING

For target q, one example is

$$\begin{aligned} \mathcal{E} &= \left(q, c, \{(\boldsymbol{x}_i, s_i, o_i, v_i)\}_{i=1}^{M}, \mathcal{A}_q\right), \quad \boldsymbol{y} \in \mathcal{Y}_q, \\ \mathcal{H}_{q,r} \subseteq \mathcal{H}_q &:= \{i \mid o_i = 1,\ \mathrm{elig}(s_i; q, c) = 1,\ v_i = \mathrm{valid}\}. \end{aligned} \tag{1}$$

Here $\boldsymbol{x}_i$ and s_i denote an observation and its source, o_i records availability, and v_i is an independently supplied validity state. The context c contains only protocol-permitted information. The versioned specification $\mathcal{A}_q$ is fixed before evaluation and does not access the outcome $\boldsymbol{y}$. It states which sources may authorize each relation r, under which conditions, and how invalid, missing, or conflicting evidence changes that authorization. Validity is distinct from estimated quality: quality may modulate how admitted messages are weighted, but it cannot create an authority relation that the specification does not permit.

Authority is therefore relation-specific, not a permanent ranking of modalities. Membership in $\mathcal{H}_q$ admits a source for the current target, whereas $\mathcal{H}_{q,r}$ identifies the sources authorized to constrain relation r. A source may be authorized for one relation and auxiliary for another, and a relation may depend jointly on several sources. The output domain is $\mathcal{D}_q = \Delta^{K-1}$ for single-label prediction and $\mathcal{D}_q = [0,1]^K$ for multilabel prediction. A valid fusion rule must satisfy two distinct invariants: perturbing unauthorized auxiliary evidence must not change a protected representation, and every returned prediction must satisfy all activated relations. Appendix A gives a minimal construction showing that observations and marginal reliability alone need not identify the relevant authority relation.

3.3 AUTHORITY SPECIFICATION COMPILER

ASC converts a declarative policy into objects that can be checked in the computational graph and output space:

$$\begin{aligned} (\mathcal{G}_q, \boldsymbol{B}_q, \boldsymbol{b}_q, \sigma) &= \mathrm{ASC}(\mathcal{A}_q; q, c, \boldsymbol{o}, \boldsymbol{v}, \boldsymbol{\tau}), \\ \mathcal{C}_q &:= \{\boldsymbol{p} \in \mathcal{D}_q \mid \boldsymbol{B}_q \boldsymbol{p} \succeq \boldsymbol{b}_q\}. \end{aligned} \tag{2}$$

The directed graph $\mathcal{G}_q$ specifies admissible information dependencies. Each active row of $\boldsymbol{B}_q \boldsymbol{p} \succeq \boldsymbol{b}_q$ encodes a permitted decision relation, and their intersection defines the authority-consistent feasible set $\mathcal{C}_q$. A row may express an ordering such as $p_a - p_b \geq \delta$ or a bound on grouped probability mass, but numerical margins are used only when supplied by the protocol. The status σ is constrained, unconstrained, conflict, or no-input. Valid and available evidence activates its authorized relations. Invalid or missing evidence withdraws them; evidence with unknown validity remains available as auxiliary information but cannot activate a hard constraint. Consequently, the same specification can compile to different graphs and feasible sets as evidence states change. Inconsistent constraints and complete absence of usable inputs produce non-predictive states.

ASC does not learn authority. Its contract is that, for fixed policy and evidence states, $\mathcal{G}_q$, $\mathcal{C}_q$, and σ depend only on declared sources, statements, availability, validity, and permitted context, never on $\boldsymbol{y}$ or an auxiliary model's predicted label. Task-specific compilation rules and pseudocode appear in Appendix A.

3.4 Authority-Protected Graph Mixer

Each source is first encoded as $\boldsymbol{h}_i^0 = E_i(\boldsymbol{x}_i)$. Our implementation summarizes token sets with mean-augmented SPD moments and operates in Log-Euclidean coordinates; Appendix A gives the construction and its Euclidean counterpart. For relation r, APGM restricts every update to parents permitted by $\mathcal{G}_q$:

$$\begin{aligned} \boldsymbol{m}_{i,r}^{\ell} &= \sum_{j \in \mathrm{Pa}_{q,r}(i)} \alpha_{ij,r}^{\ell} \boldsymbol{V}_{ij,r}^{\ell} \boldsymbol{h}_{j,r}^{\ell}, \\ \boldsymbol{h}_{i,r}^{\ell+1} &= \phi_{i,r}^{\ell}(\boldsymbol{h}_{i,r}^{\ell}, \boldsymbol{m}_{i,r}^{\ell}). \end{aligned} \tag{3}$$

The relation index retains multiple simultaneous authorization patterns instead of collapsing them into one global modality graph. Protected nodes may receive messages only from the sources authorized for that relation and from immutable support nodes. Auxiliary nodes may read protected states, but no multi-hop path, shared normalization, gate, cache, or support node may relay auxiliary information back into them. Attention coefficients are normalized only within the allowed neighborhood. They allocate a fixed communication budget after authorization has been determined; edge allocation and authority are consequently separate variables. Learned attention can choose among permitted senders, but cannot reopen a prohibited dependency.

The protected mixing rule yields fixed-parameter non-interference when the specification and authorized inputs are fixed, the protected subgraph is closed under computational ancestry, and no bypass reads unauthorized inputs. Its proof is an induction over graph layers. Stop-gradient alone does not satisfy the premise because it blocks optimization signals, not forward information. Appendix A provides the proof and a bypass counterexample.

3.5 Authority-Consistent Projection and Learning

APGM protects designated intermediate states, but an unrestricted prediction head could still combine auxiliary states into an inadmissible output. The head therefore first produces $\boldsymbol{p}_0$, after which ACP returns the closest feasible prediction:

$$\boldsymbol{p}^{\star} := \Pi_{\mathcal{C}_q}(\boldsymbol{p}_0) = \operatorname*{arg\,min}_{\boldsymbol{p} \in \mathcal{C}_q} \frac{1}{2} \|\boldsymbol{p} - \boldsymbol{p}_0\|_2^2. \tag{4}$$

Let $d_K(\boldsymbol{a}, \boldsymbol{b}) := K^{-1}\|\boldsymbol{a} - \boldsymbol{b}\|_2^2$; its first occurrence below is the dimension-normalized Brier loss. The training objective is

$$\mathcal{L}_{\mathrm{ACF}} = d_K(\boldsymbol{p}^{\star}, \boldsymbol{y}) + \lambda_{\mathrm{proj}} d_K(\boldsymbol{p}^{\star}, \boldsymbol{p}_0), \qquad \lambda_{\mathrm{proj}} = 0.1. \tag{5}$$

Together with ASC's feasibility check and APGM's dependency restrictions, ACP implements an authority-consistent minimum-intervention principle: it enforces the activated relations while preserving the model's remaining preferences. Auxiliary evidence can therefore resolve unconstrained degrees of freedom. If no relation is active, ACF returns $\boldsymbol{p}_0$; conflict or absence of usable input causes abstention, while numerical failure is recorded separately. The first term of Eq. 5 trains the returned probability and the second discourages systematic reliance on large corrections.

For a fixed nonempty closed convex $\mathcal{C}_q$, the projection is unique and non-expansive. Writing $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{Y} \mid \mathcal{I}]$, if $\boldsymbol{\mu} \in \mathcal{C}_q$, conditional dimension-normalized Brier risk improves by at least $K^{-1}\|\boldsymbol{p}_0 - \boldsymbol{p}^{\star}\|_2^2$; under misspecification, its excess is at most $K^{-1}\,\mathrm{dist}^2(\boldsymbol{\mu}, \mathcal{C}_q)$. Feasibility certificates, optimality conditions, task-specific closed forms, proofs, and failure cases are given in Appendix A. Algorithm 1 summarizes the end-to-end procedure: authority compilation remains label-free, while the projected Brier objective jointly updates the encoders, protected graph mixer, and prediction head.

Table 1: **Quantitative results on UTSW MRI-to-semantic retrieval.** Higher values are better.

Method	mAP ↑	R@1 ↑	MRR ↑
SPD-UB	**0.7267**	**0.7605**	**0.8469**
SPD uniform graph	0.7256	0.7575	0.8447
SPD learned graph	0.7054	0.7527	0.8462
Euclidean learned graph	0.6405	0.7413	0.8429
Latent concatenation	0.6298	0.7171	0.8313

Table 2: **Module contributions on UTSW.** Each row reports the mAP gain after adding the corresponding component.

Component	Without	With	Gain
SPD representation	0.6405	0.7054	+6.485
Graph communication	0.7007	0.7256	+2.482
Upper-level communication	0.7007	0.7267	+2.593
Semantic anchors	0.6867	0.7054	+1.867
Uniform edge allocation	0.7054	0.7256	+2.018

4 EXPERIMENTS

We report the main predictive, architectural, and mechanism results in Tables 1–3. We evaluate multimodal retrieval on UTSW and examine authority mechanisms using the *Authority Identification Suite* (AIS). UTSW is evaluated by mean average precision (mAP), recall at rank one (R@1), and mean reciprocal rank (MRR). AIS contains two diagnostic tasks: S1 evaluates target-conditioned class relations, and S2 evaluates relation-specific probability orders. Predictive utility is measured by Brier score, decision authority by authority-violation rate (AVR), and computational protection by protected-state shift. Appendix B provides the complete experimental record, including task construction, information-matched baselines, seed-level statistics, intervention grids, efficiency, and reproducibility audits. The AIS protocol and baseline information are detailed in Appendix B.1 and Table B.2.

4.1 QUANTITATIVE RESULTS

Table 1 reports the quantitative results on UTSW MRI-to-semantic retrieval. SPD-UB achieves an mAP of 0.7267, R@1 of 0.7605, and MRR of 0.8469. Compared with latent concatenation, it improves mAP by 9.687 points, R@1 by 4.34 points, and MRR by 1.56 points. It also improves mAP by 8.613 points over Euclidean graph fusion. These results show that combining geometric representation with structured communication provides an effective multimodal fusion backbone. Complete confirmation results, variability across runs, and the separately reported screening baselines are available in Appendix B.6, Table B.11, and Figure 7.

4.2 MODULE ABLATION

Table 2 isolates the contribution of each fusion component. SPD representation contributes 6.485 mAP points, while graph communication contributes 2.482 points. Upper-level communication adds 2.593 points, and semantic anchors add 1.867 points. Uniform edge allocation further improves mAP by 2.018 points. The largest gain comes from representation geometry, followed by cross-modal communication and semantic support. Appendix Table B.12 reports the complete paired contrasts and graph interventions. Target-specific anchor effects and learned-edge stability are provided in Table B.13, while Figure 8 visualizes graph co-adaptation, subgroup behavior, and edge reproducibility.

4.3 MECHANISM ANALYSIS

The Authority Identification Suite separates predictive utility, reliability, computational protection, and decision authority. Table 3 shows that ACF preserves the predictive utility of information-

Table 3: **Mechanism analysis on the Authority Identification Suite.** The experiments separate predictive utility, reliability, computational protection, and decision authority.

Mechanism	Measurement	S1	S2
Predictive utility	Absolute Brier difference from information-matched learned fusion	5.5×10^{-5}	6.6×10^{-5}
Reliability	Brier improvement as auxiliary reliability increases from 0.60 to 0.95	N/A	5.8×10^{-3}
Decision authority	AVR before and after authority-consistent projection	$1.0 \rightarrow 0$	$0 \rightarrow 0$
Computational protection	Protected-state shift before and after protected graph computation	$0.1711 \rightarrow 0$	$0.1247 \rightarrow 0$
Policy specification	Brier improvement of the correct policy over an incorrect policy	1.888×10^{-1}	8.39×10^{-2}
Relation specification	Brier improvement of the correct relation over an incorrect relation	3.391×10^{-1}	1.681×10^{-1}

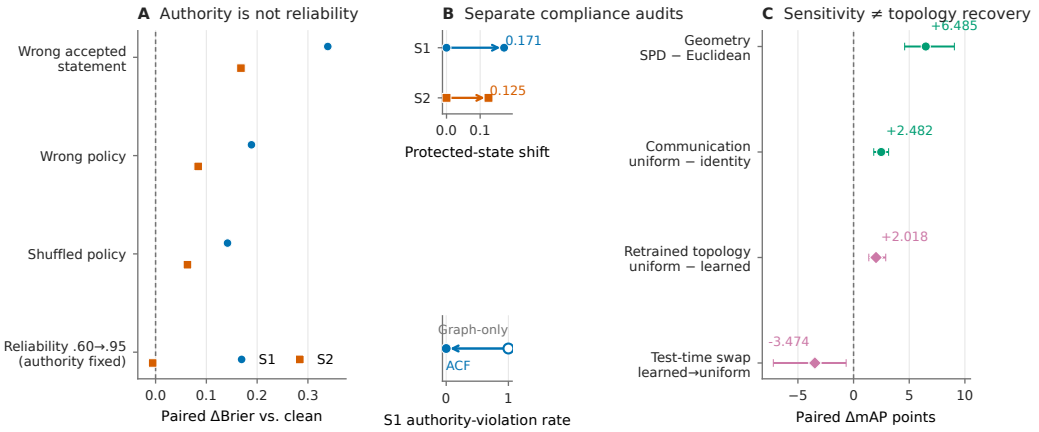


Figure 3: **Predictive utility, reliability, authority, and graph structure are identified through separate interventions.** (A) Policy and reliability interventions produce distinct changes in prediction risk. (B) Protected graph computation prevents unauthorized state changes, while projection enforces the activated decision relation. (C) Representation geometry, communication, and edge allocation make separate contributions to UTSW MRI-to-semantic retrieval.

matched graph fusion, with Brier differences below 7.0×10^{-5}. Increasing auxiliary reliability improves Brier by 5.8×10^{-3} without changing the activated authority relation. Authority-consistent projection reduces the S1 violation rate from 1 to 0, while protected graph computation eliminates auxiliary-induced state shifts.

Correct policies improve Brier by 1.888×10^{-1} on S1 and 8.39×10^{-2} on S2. Correct accepted relations provide further improvements of 3.391×10^{-1} and 1.681×10^{-1}, respectively. Complete predictive results for all information-matched methods are reported in Appendix Table B.3, with paired ACF comparisons in Table B.4. Policy, relation, reliability, and conflict interventions are expanded in Tables B.5 and B.7. Computational protection, failure states, and abstention behavior are audited in Tables B.8 and B.6, and visualized in Figure 6.

Figure 3 connects the three levels of evidence reported in the main text. Panel A distinguishes authority specification from source reliability. Panel B reports computational protection and decision compliance. Panel C separates representation geometry, graph communication, and edge allocation. The corresponding run-level distributions and intervention results are provided in Figures 4, 5, and 7. All reported configurations, result files, and recomputed metrics are summarized in the reproducibility audit in Appendix B.7 and Table B.14.

5 Discussion and Conclusion

These results support treating evidential authority as an explicit task variable rather than inferring it from predictive utility, reliability, or learned topology. ACF matches information-matched graph fusion within 7.0×10^{-5} Brier while enforcing decision constraints and protecting authorized states. Incorrect policies increase Brier by 8.4×10^{-2} to 1.89×10^{-1}, showing that authority helps only when correctly specified. UTSW attribution separates geometry, communication, and edge allocation through matched retraining and graph interventions, preventing graph dependence from being mistaken for topology or authority recovery. Our guarantees require valid affine constraints, closed convex feasible sets, and protected computation without bypasses; controlled tasks and single-center retrieval do not constitute clinical validation. Authority should therefore be specified, intervened on, and audited as part of multimodal decision design.

References

Brandon Amos and J. Zico Kolter. Optnet: Differentiable optimization as a layer in neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 136–145, 2017.

Henri Arno and Thomas Demeester. Annotation-assisted learning of treatment policies from multimodal electronic health records. In *Proceedings of the 11th Machine Learning for Healthcare Conference*, volume 340 of *Proceedings of Machine Learning Research*, pp. 77–110. PMLR, 2026.

Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 29(1):328–347, 2007. doi: 10.1137/050637996.

Alexander Bakumenko, Dane Hudson Smith, and Janine Hoelscher. Early-horizon multimodal ICU mortality prediction without retraining. In *Proceedings of the 11th Machine Learning for Healthcare Conference*, volume 340 of *Proceedings of Machine Learning Research*, pp. 214–240. PMLR, 2026.

Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2 edition, 2017. doi: 10.1007/978-3-319-48311-5.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Alexander Campbell, Antonio Giuliano Zippo, Luca Passamonti, Nicola Toschi, and Pietro Lio. Dyndepnet: Learning time-varying dependency structures from fmri data via dynamic graph structure learning. In *ICML Workshop on Interpretable Machine Learning in Healthcare*, 2023.

Hei Ting Una Chan, Chenwei Wu, Xueshen Liu, Zesen Zhao, Boyuan Zheng, Luis Filipe Nakayama, Michael G. Morley, Liyue Shen, Jiasi Chen, and Z. Morley Mao. A cloud–edge system for multimodal clinical screening in resource-constrained rural settings. In *Proceedings of the 11th Machine Learning for Healthcare Conference*, volume 340 of *Proceedings of Machine Learning Research*, pp. 283–313. PMLR, 2026.

Zixian Gao, Xun Jiang, Xing Xu, Fumin Shen, Yujie Li, and Heng Tao Shen. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26876–26885, 2024.

Linxiao Gong, Yang Liu, Lianlong Sun, Yulai Bi, Jing Liu, and Xiaoguang Zhu. Embracing aleatoric uncertainty in medical multimodal learning with missing modalities. *arXiv preprint arXiv:2601.21950*, 2026.

Kshitij Goyal, Sebastijan Dumancic, and Hendrik Blockeel. Deepsade: Learning neural networks that guarantee domain constraint satisfaction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 12199–12207, 2024. doi: 10.1609/aaai.v38i11.29109.

GRADE Working Group. Grading quality of evidence and strength of recommendations. *BMJ*, 328 (7454):1490, 2004. doi: 10.1136/bmj.328.7454.1490.

Gordon Guyatt, Per Olav Vandvik, Alfonso Iorio, Arnav Agarwal, Liang Yao, Prashanti Eachempati, Linan Zeng, Derek K. Chu, Rohan D'Souza, Thomas Agoritsas, M. Hassan Murad, Stefan Schandelmaier, Jamie Rylance, Benjamin Djulbegovic, Victor M. Montori, Monica Hultcrantz, and Romina Brignardello-Petersen. Core GRADE 7: Principles for moving from evidence to recommendations and decisions. *BMJ*, 389:e083867, 2025. doi: 10.1136/bmj-2024-083867.

Ji Hun Ham, Daniel D. Lee, and Lawrence K. Saul. Semisupervised alignment of manifolds. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pp. 120–127, 2005.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. In *International Conference on Learning Representations*, 2021.

Zongliang Ji, Cian Eastwood, Anna Goldenberg, Paul Pu Liang, Jason Hartford, Rahul G. Krishnan, and Emmanuel Noutahi. Unpaired multimodal learning for biological datasets. In *Proceedings of the 9th International Conference on Medical Imaging with Deep Learning*, volume 315 of *Proceedings of Machine Learning Research*, pp. 1840–1868. PMLR, 2026.

Seowung Leem, Lin Gu, Chenyu You, Kuang Gong, and Ruogu Fang. REVEAL: Multimodal vision–language alignment of retinal morphometry and clinical risks for incident AD and dementia prediction. In *Proceedings of the 9th International Conference on Medical Imaging with Deep Learning*, volume 315 of *Proceedings of Machine Learning Research*, pp. 1869–1889. PMLR, 2026.

Xuanchen Li, Yuheng Lu, Chenrui Cui, Tianrui Wang, Zikang Huang, Yu Jiang, Long Zhou, Longbiao Wang, and Jianwu Dang. Separate first, fuse later: Mitigating cross-modal interference in audio-visual llms reasoning with modality-specific chain-of-thought. *arXiv preprint arXiv:2605.09906*, 2026.

Xinyan Liang, Shijie Wang, Yuhua Qian, Qian Guo, Liang Du, Bingbing Jiang, and Tingjin Luo. Trusted multi-view classification with expert knowledge constraints. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 2025.

Alessandro Manenti, Daniele Zambon, and Cesare Alippi. Learning latent graph structures and their uncertainty. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 2025.

Alireza Moayedikia. Conformal fusion under missing modalities. *arXiv preprint arXiv:2608.07183*, 2026.

Nusrat Binta Nizam, Fengbei Liu, Sunwoo Kwak, Ilan Richter, Jayant K. Raikhelkar, Ashley Beecy, Nir Uriel, Deborah Estrin, and Mert R. Sabuncu. X-cardia: Phenotype-guided cross-modal alignment for opportunistic cardiac screening on routine chest CT. In *Proceedings of the 9th International Conference on Medical Imaging with Deep Learning*, volume 315 of *Proceedings of Machine Learning Research*, pp. 1740–1767. PMLR, 2026.

Vincent Ochs, Christoph Kuemmerli, Florentin Bieder, Julia Wolleb, Jo"el L. Lavanchy, Julia Ruppel, Jan Liechti, Stephanie Taha-Mehlitz, Christian Andreas Nebiker, Beat M"uller, Giuseppe Kito Fusai, Joerg-Matthias Pollok, Anas Taha, Philippe C. Cattin, and Sebastian Staubli. Multimodal assessment of pancreatic cancer resectability using deep learning. In *Proceedings of the 9th International Conference on Medical Imaging with Deep Learning*, volume 315 of *Proceedings of Machine Learning Research*, pp. 88–106. PMLR, 2026.

Petra Poklukar, Miguel Vasco, Hang Yin, Francisco S. Melo, Ana Paiva, and Danica Kragic. Geometric multimodal contrastive representation learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 17782–17800, 2022.

Chongyu Qu, Zhengyi Lu, Yuxiang Lai, Thomas Z. Li, Junchao Zhu, Junlin Guo, Juming Xiong, Yanfan Zhu, Yuechen Yang, Allen J. Luna, Kim L. Sandler, Bennett A. Landman, and Yuankai Huo. Adafuse: Adaptive multimodal fusion for lung cancer risk prediction via reinforcement

learning. In *Proceedings of the 9th International Conference on Medical Imaging with Deep Learning*, volume 315 of *Proceedings of Machine Learning Research*, pp. 2582–2601. PMLR, 2026.

Abhinav Sagar. TAFIE: Transformer-assisted fusion with integrated entropy attention for multimodal medical imaging. In *Proceedings of the Second AAAI Bridge Program on AI for Medicine and Healthcare*, volume 317 of *Proceedings of Machine Learning Research*, pp. 229–238. PMLR, 2026.

Nafis Fuad Shahid, Maroof Ahmed, Md Akib Haider, Saidur Rahman Sagor, Aashnan Rahman, and Md Azam Hossain. Probabilistic feature imputation and uncertainty-aware multimodal federated aggregation. In *Proceedings of the 9th International Conference on Medical Imaging with Deep Learning*, volume 315 of *Proceedings of Machine Learning Research*, pp. 3593–3607. PMLR, 2026.

Hang Shi, Changxi Chi, Peng Wan, and Daoqiang Zhang. Multi-modal topology-embedded graph learning for spatially resolved genes prediction from pathology images with prior gene similarity information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Tao Tang, Guangrun Wang, Yixing Lao, Peng Chen, Jie Liu, Liang Lin, Kaicheng Yu, and Xiaodan Liang. Alignmif: Geometry-aligned multimodal implicit field for lidar-camera joint synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21740–21751, 2024.

Megan Tjandrasuwita, Chanakya Ekbote, Ziyin Liu, and Paul Pu Liang. Understanding the emergence of multimodal representation alignment. *arXiv preprint arXiv:2502.16282*, 2025.

Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multimodal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15878–15887, 2023.

Yucheng Xing, Hailan Mo, Zi Wang, Ling Huang, and Mengling Feng. Evidential fusion network for multimodal survival prediction under missing modalities. *arXiv preprint arXiv:2606.20757*, 2026.

Minji Yoon, Jing Yu Koh, Bryan Hooi, and Ruslan Salakhutdinov. Multimodal graph learning for generative tasks. *arXiv preprint arXiv:2310.07478*, 2023.

Linda Trinkaus Zagzebski. *Epistemic Authority: A Theory of Trust, Authority, and Autonomy in Belief*. Oxford University Press, 2012. doi: 10.1093/acprof:oso/9780199936472.001.0001.

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 *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 41753–41769, 2023.

Yuan Zhang, Yutong Xie, Hu Wang, Jodie C. Avery, M. Louise Hull, and Gustavo Carneiro. A novel perspective for multi-modal multi-label skin lesion classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3549–3558, 2025. doi: 10.1109/WACV61041.2025.00350.

A Authority-Constrained Minimum-Intervention Principle

This appendix formalizes ACF as a constrained learning principle rather than as an unconstrained predictor followed by an ad hoc rule correction. Given a label-free authority specification, ACF first determines whether the activated relations admit a valid decision, restricts the computational dependencies through which protected evidence may be updated, and finally returns the feasible prediction that minimally changes the unconstrained model output. The resulting chain is

$$\text{specification feasibility} \longrightarrow \text{admissible computation} \longrightarrow \text{minimum-intervention decision}.$$

The convex-analytic ingredients are standard (Boyd & Vandenberghe, 2004; Bauschke & Combettes, 2017); our use of them is to connect an externally specified authority relation to separately auditable computational and decision constraints. Differentiable quadratic programming itself is not claimed as a contribution (Amos & Kolter, 2017).

A.1 Authority-Constrained Decision Problem

Consider an evidence record

$$\mathcal{E} = \left(q, c, \{(\boldsymbol{x}_i, s_i, o_i, v_i)\}_{i=1}^{M}, \mathcal{A}_q\right),$$

where q is the target, c is admissible context, o_i and v_i respectively denote availability and independently verified validity, and $\mathcal{A}_q$ is a versioned authority specification fixed before observing the target label. The compiler is label-free: it may use $\mathcal{A}_q, q, c, \boldsymbol{o}, \boldsymbol{v}$, and authorized evidence statements, but neither the training or test label nor a prediction from an auxiliary source.

We write both single-label and multilabel output domains as the convex polytope

$$\mathcal{D}_q = \left\{\boldsymbol{p} \in \mathbb{R}^K : \boldsymbol{A}_D \boldsymbol{p} \le \boldsymbol{d}_D,\ \boldsymbol{E}_D \boldsymbol{p} = \boldsymbol{e}_D\right\}. \tag{A.1}$$

For a single-label problem, $\mathcal{D}_q = \Delta^{K-1} = \{\boldsymbol{p} \ge 0 : \mathbf{1}^\top \boldsymbol{p} = 1\}$. For a multilabel probability problem, $\mathcal{D}_q = [0,1]^K$, with no sum-to-one constraint.

Let $\mathcal{R}(\mathcal{E})$ denote the relations activated by available and verified authority sources. Inactive relations are removed rather than represented by artificial zero rows. The active authority-feasible set is

$$\mathcal{C}(\mathcal{E}) = \{\boldsymbol{p} \in \mathcal{D}_q : \boldsymbol{B}_{\mathcal{E}} \boldsymbol{p} \ge \boldsymbol{b}_{\mathcal{E}}\}. \tag{A.2}$$

Each row may encode, for example, an order relation $p_a - p_b \ge \delta$, a lower or upper bound on a group of outcomes, or the exclusion of inadmissible classes. Any nonzero margin δ must be part of the external task specification or a preregistered design parameter. It is not inferred from attention weights, validation performance, or modality reliability.

A.2 Geometry and Feasibility

Lemma A.1 (Geometry of the authority-feasible set)**.** *Suppose that $\mathcal{D}_q$ is a nonempty compact convex polytope and that every activated authority relation is affine. Then $\mathcal{C}(\mathcal{E})$ is closed and convex. If it is nonempty, it is also compact.*

Proof. For the r-th row of $(\boldsymbol{B}_{\mathcal{E}}, \boldsymbol{b}_{\mathcal{E}})$, define

$$\mathcal{H}_r = \{\boldsymbol{p} : (\boldsymbol{B}_{\mathcal{E}})_{r:} \boldsymbol{p} \ge (\boldsymbol{b}_{\mathcal{E}})_r\}.$$

Because the defining map is affine, $\mathcal{H}_r$ is a closed convex halfspace. Equation A.2 gives

$$\mathcal{C}(\mathcal{E}) = \mathcal{D}_q \cap \bigcap_{r \in \mathcal{R}(\mathcal{E})} \mathcal{H}_r.$$

A finite intersection of closed convex sets is closed and convex. Moreover, $\mathcal{C}(\mathcal{E}) \subseteq \mathcal{D}_q$, so it is bounded. A nonempty closed and bounded subset of finite-dimensional Euclidean space is compact. □

Lemma A.1 does not guarantee nonemptiness. Treating nonemptiness as an unchecked assumption would conflate an inconsistent authority specification with a numerical solver failure. We therefore define the Phase-I problem

$$\begin{aligned} s^\star(\mathcal{E}) := \min_{\boldsymbol{p}, \boldsymbol{\xi}} \quad & \mathbf{1}^\top \boldsymbol{\xi} \\ \text{s.t.} \quad & \boldsymbol{p} \in \mathcal{D}_q, \qquad \boldsymbol{\xi} \ge 0, \\ & \boldsymbol{B}_{\mathcal{E}} \boldsymbol{p} + \boldsymbol{\xi} \ge \boldsymbol{b}_{\mathcal{E}}. \end{aligned} \tag{A.3}$$

Proposition A.2 (Feasibility certificate)**.** *Under the conditions of Lemma A.1,*

$$\mathcal{C}(\mathcal{E}) \neq \varnothing \quad \Longleftrightarrow \quad s^\star(\mathcal{E}) = 0.$$

Proof. If $\bar{\boldsymbol{p}} \in \mathcal{C}(\mathcal{E})$, then $(\bar{\boldsymbol{p}}, \mathbf{0})$ is feasible for equation A.3, so $s^\star \le 0$. Since $\boldsymbol{\xi} \ge 0$, the objective is nonnegative and hence $s^\star = 0$.

Conversely, suppose $s^\star = 0$. The optimum is attained: choose any $\boldsymbol{p} \in \mathcal{D}_q$ and sufficiently large nonnegative slack to obtain a finite feasible value, then restrict attention to the corresponding compact objective sublevel set. At an optimum with $\mathbf{1}^\top \boldsymbol{\xi}^\star = 0$, nonnegativity implies $\boldsymbol{\xi}^\star = \mathbf{0}$. Therefore $\boldsymbol{B}_{\mathcal{E}} \boldsymbol{p}^\star \ge \boldsymbol{b}_{\mathcal{E}}$, and $\boldsymbol{p}^\star \in \mathcal{C}(\mathcal{E})$. □

A dual certificate can also be obtained from Farkas' lemma. Stack the domain and authority inequalities as

$$\bar{\boldsymbol{A}} = \begin{bmatrix} \boldsymbol{A}_D \\ -\boldsymbol{B}_{\mathcal{E}} \end{bmatrix}, \qquad \bar{\boldsymbol{d}} = \begin{bmatrix} \boldsymbol{d}_D \\ -\boldsymbol{b}_{\mathcal{E}} \end{bmatrix}.$$

Exactly one of the following systems is feasible:

$$\bar{\boldsymbol{A}} \boldsymbol{p} \le \bar{\boldsymbol{d}}, \qquad \boldsymbol{E}_D \boldsymbol{p} = \boldsymbol{e}_D,$$

or

$$\bar{\boldsymbol{A}}^\top \boldsymbol{\lambda} + \boldsymbol{E}_D^\top \boldsymbol{\nu} = 0, \qquad \boldsymbol{\lambda} \ge 0, \qquad \bar{\boldsymbol{d}}^\top \boldsymbol{\lambda} + \boldsymbol{e}_D^\top \boldsymbol{\nu} < 0. \tag{A.4}$$

Equation A.4 is an explicit inconsistency certificate. In ACF, a mathematically infeasible compilation receives the `conflict` status and produces no returned prediction. Failure of a numerical method to establish feasibility is instead recorded as `solver_failure`. The two statuses have different meanings and are never merged in evaluation.

A.3 Minimum-Intervention Decision Principle

Let $\boldsymbol{p}_0 \in \mathcal{D}_q$ be the unconstrained probability prediction produced by the learned fusion network. Whenever $\mathcal{C}(\mathcal{E}) \neq \varnothing$, the authority-consistent decision is defined by

$$\boldsymbol{p}^\star = P_{\mathcal{C}(\mathcal{E})}(\boldsymbol{p}_0) := \arg\min_{\boldsymbol{p} \in \mathcal{C}(\mathcal{E})} \frac{1}{2} \|\boldsymbol{p} - \boldsymbol{p}_0\|_2^2. \tag{A.5}$$

Multiplication of the objective by $1/K$ does not change its minimizer. We retain that factor later when relating the construction to the dimension-normalized Brier score.

Equation A.5 is the *authority-consistent minimum-intervention principle*: enforce every valid activated relation, but introduce no larger Euclidean change than required for feasibility. Auxiliary evidence may therefore affect any degree of freedom remaining inside $\mathcal{C}(\mathcal{E})$, but it cannot move the returned prediction outside that set.

Theorem A.3 (Existence, uniqueness, and minimal intervention)**.** *If $\mathcal{C}(\mathcal{E})$ is nonempty, closed, and convex, then equation A.5 has a unique solution. Moreover,*

$$\boldsymbol{p}_0 \in \mathcal{C}(\mathcal{E}) \quad \Longrightarrow \quad \boldsymbol{p}^\star = \boldsymbol{p}_0,$$

and, for every $\boldsymbol{p} \in \mathcal{C}(\mathcal{E})$,

$$\|\boldsymbol{p}^\star - \boldsymbol{p}_0\|_2 \le \|\boldsymbol{p} - \boldsymbol{p}_0\|_2.$$

Proof. Existence follows because the continuous objective in equation A.5 attains its minimum on the nonempty compact set established in Lemma A.1. The squared Euclidean norm is strictly convex in $\boldsymbol{p}$, so it cannot have two distinct minimizers on a convex set. The last inequality is the definition of optimality. If $\boldsymbol{p}_0 \in \mathcal{C}(\mathcal{E})$, it attains objective value zero, which is the global lower bound; uniqueness then gives $\boldsymbol{p}^\star = \boldsymbol{p}_0$. □

The Euclidean metric is not presented as the only possible notion of intervention. It is selected because it matches the Brier geometry used for training and evaluation and yields the stability and conditional-risk results below. Replacing it with a different divergence requires new guarantees.

A.4 Optimality Conditions and Registered Closed Forms

Theorem A.4 (Projection and KKT characterizations)**.** *For a nonempty closed convex set $\mathcal{C}$, a point $\boldsymbol{p}^\star \in \mathcal{C}$ solves equation A.5 if and only if*

$$\langle \boldsymbol{p}_0 - \boldsymbol{p}^\star,\, \boldsymbol{p} - \boldsymbol{p}^\star \rangle \le 0, \qquad \forall \boldsymbol{p} \in \mathcal{C}. \tag{A.6}$$

For the polyhedral set in equation A.2, this is equivalent to the existence of multipliers $\boldsymbol{\lambda}^\star \ge 0$ and $\boldsymbol{\nu}^\star$ satisfying

$$\begin{aligned} &\boldsymbol{p}^\star - \boldsymbol{p}_0 + \bar{\boldsymbol{A}}^\top \boldsymbol{\lambda}^\star + \boldsymbol{E}_D^\top \boldsymbol{\nu}^\star = 0, \\ &\bar{\boldsymbol{A}}\boldsymbol{p}^\star \le \bar{\boldsymbol{d}}, \qquad \boldsymbol{E}_D \boldsymbol{p}^\star = \boldsymbol{e}_D, \\ &\boldsymbol{\lambda}^\star \odot (\bar{\boldsymbol{A}}\boldsymbol{p}^\star - \bar{\boldsymbol{d}}) = 0. \end{aligned} \tag{A.7}$$

Proof. Assume first that $\boldsymbol{p}^\star$ minimizes equation A.5. For any $\boldsymbol{p} \in \mathcal{C}$, convexity implies that $\boldsymbol{p}^\star + t(\boldsymbol{p} - \boldsymbol{p}^\star) \in \mathcal{C}$ for $t \in [0, 1]$. The one-sided directional derivative of the objective at $t = 0$ must be nonnegative:

$$\langle \boldsymbol{p}^\star - \boldsymbol{p}_0,\, \boldsymbol{p} - \boldsymbol{p}^\star \rangle \ge 0,$$

which is equation A.6.

Conversely, expand the objective at any $\boldsymbol{p} \in \mathcal{C}$:

$$\|\boldsymbol{p} - \boldsymbol{p}_0\|_2^2 = \|\boldsymbol{p} - \boldsymbol{p}^\star\|_2^2 + \|\boldsymbol{p}^\star - \boldsymbol{p}_0\|_2^2 + 2\langle \boldsymbol{p} - \boldsymbol{p}^\star,\, \boldsymbol{p}^\star - \boldsymbol{p}_0 \rangle.$$

The last term is nonnegative by equation A.6, proving that $\boldsymbol{p}^\star$ is optimal. The KKT system follows from the normal-cone representation of a polyhedron. Conversely, its stationarity, primal feasibility, dual feasibility, and complementary slackness conditions imply equation A.6, and hence optimality. □

The multipliers in equation A.7 quantify which active inequalities contribute locally to the correction from $\boldsymbol{p}_0$ to $\boldsymbol{p}^\star$. They need not be unique even when $\boldsymbol{p}^\star$ is unique, and their magnitude is not an evidential-authority score.

S1: restricted simplex. Let $\mathcal{K} \subseteq \{1, \dots, K\}$ be the classes admitted by an activated S1 relation. The feasible set is the corresponding simplex face:

$$\mathcal{C}_\mathcal{K} = \left\{ \boldsymbol{p} \in \Delta^{K-1} : p_k = 0 \text{ for } k \notin \mathcal{K} \right\}.$$

Corollary A.5 (Restricted-simplex projection)**.** *The unique projection onto $\mathcal{C}_\mathcal{K}$ is*

$$p_k^\star = \begin{cases} [p_{0,k} - \tau]_+, & k \in \mathcal{K}, \\ 0, & k \notin \mathcal{K}, \end{cases} \qquad \sum_{k \in \mathcal{K}} [p_{0,k} - \tau]_+ = 1, \tag{A.8}$$

where the scalar threshold τ is uniquely determined by the normalization equation.

Proof. Apply the KKT conditions to the equality $\sum_{k \in \mathcal{K}} p_k = 1$, nonnegativity, and the fixed zero coordinates outside $\mathcal{K}$. Stationarity on admitted coordinates gives $p_k^\star = p_{0,k} - \tau + \gamma_k$, where $\gamma_k \ge 0$ is the multiplier for nonnegativity. Complementary slackness yields the threshold form in equation A.8. The left-hand side of its normalization equation is continuous and strictly decreasing over its nonconstant range, so the required threshold is unique. □

S2: disjoint order constraints. Consider an active relation $p_a \ge p_b$ with $(p_{0,a}, p_{0,b}) \in [0, 1]^2$.

Corollary A.6 (Projection onto a pairwise order)**.** *The projection is*

$$(p_a^\star, p_b^\star) = \begin{cases} (p_{0,a}, p_{0,b}), & p_{0,a} \ge p_{0,b}, \\ \left(\frac{p_{0,a}+p_{0,b}}{2}, \frac{p_{0,a}+p_{0,b}}{2} \right), & p_{0,a} < p_{0,b}. \end{cases}$$

The reverse relation is obtained by exchanging a and b. If the active relations involve disjoint coordinate pairs, the full projection is the concatenation of these pairwise projections and the unchanged unconstrained coordinates.

Proof. If the relation is already satisfied, Theorem A.3 gives the unchanged point. Otherwise the closest point lies on the boundary $p_a = p_b = t$. Minimizing

$$(t - p_{0,a})^2 + (t - p_{0,b})^2$$

gives $t = (p_{0,a} + p_{0,b})/2$, which remains in $[0, 1]$. Disjoint coordinate pairs make both the objective and constraints separable. Overlapping relations do not admit this decomposition and are handled by the general quadratic program. □

These two corollaries justify the registered analytic solvers used in the controlled tasks. They are exact special cases of equation A.5, not approximations to the general layer.

A.5 Stability and Relation Withdrawal

Theorem A.7 (Firm non-expansiveness)**.** *Let $\mathcal{C}$ be fixed, nonempty, closed, and convex. For any $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^K$,*

$$\|P_{\mathcal{C}}(\boldsymbol{u}) - P_{\mathcal{C}}(\boldsymbol{v})\|_2^2 \leq \langle P_{\mathcal{C}}(\boldsymbol{u}) - P_{\mathcal{C}}(\boldsymbol{v}), \boldsymbol{u} - \boldsymbol{v} \rangle . \tag{A.9}$$

Consequently,

$$\|P_{\mathcal{C}}(\boldsymbol{u}) - P_{\mathcal{C}}(\boldsymbol{v})\|_2 \leq \|\boldsymbol{u} - \boldsymbol{v}\|_2. \tag{A.10}$$

Proof. Let $\boldsymbol{a} = P_{\mathcal{C}}(\boldsymbol{u})$ and $\boldsymbol{d} = P_{\mathcal{C}}(\boldsymbol{v})$. Applying equation A.6 first with $(\boldsymbol{p}_0, \boldsymbol{p}^\star, \boldsymbol{p}) = (\boldsymbol{u}, \boldsymbol{a}, \boldsymbol{d})$ and then with $(\boldsymbol{v}, \boldsymbol{d}, \boldsymbol{a})$ gives

$$\langle \boldsymbol{u} - \boldsymbol{a}, \boldsymbol{d} - \boldsymbol{a} \rangle \leq 0, \qquad \langle \boldsymbol{v} - \boldsymbol{d}, \boldsymbol{a} - \boldsymbol{d} \rangle \leq 0.$$

Rearranging and adding yields

$$\langle \boldsymbol{u} - \boldsymbol{v}, \boldsymbol{a} - \boldsymbol{d} \rangle \geq \|\boldsymbol{a} - \boldsymbol{d}\|_2^2,$$

which is equation A.9. Cauchy–Schwarz gives equation A.10. □

This result concerns perturbations of probability vectors under a fixed compiled set. It is not a Lipschitz guarantee from images, text, or other raw inputs to predictions. If a perturbation changes validity, availability, context, or the authority specification, then it also changes $\mathcal{C}$, and Theorem A.7 cannot be applied directly.

Proposition A.8 (Withdrawal of activated relations)**.** *Suppose that missingness or invalidity only removes authority constraints and activates no fallback rule. Then*

$$\mathcal{C}_{\text{before}} \subseteq \mathcal{C}_{\text{after}}$$

and, for every $\boldsymbol{p}_0$,

$$\operatorname{dist}(\boldsymbol{p}_0, \mathcal{C}_{\text{after}}) \leq \operatorname{dist}(\boldsymbol{p}_0, \mathcal{C}_{\text{before}}).$$

Proof. Every point satisfying all constraints before withdrawal satisfies the remaining subset afterward, establishing set inclusion. The minimum distance over a superset cannot exceed the minimum over its subset. □

The proposition concerns correction magnitude only. It does not impose a coordinatewise ordering between the two projected predictions, nor does it apply when a fallback policy recompiles a different set of relations.

A.6 Conditional Brier Risk

Define the dimension-normalized squared distance

$$d_K(\boldsymbol{a}, \boldsymbol{b}) = \frac{1}{K}\|\boldsymbol{a} - \boldsymbol{b}\|_2^2.$$

Let $\mathcal{I}$ contain all information actually available to the predictor, including the observed sources, target, policy, validity, and availability state. Write

$$\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{Y} \mid \mathcal{I}], \qquad R(\boldsymbol{p} \mid \mathcal{I}) = \mathbb{E}[d_K(\boldsymbol{p}, \boldsymbol{Y}) \mid \mathcal{I}],$$

where $\boldsymbol{p}$ is $\mathcal{I}$-measurable. No independence among multilabel outcomes is required.

Lemma A.9 (Conditional squared-loss decomposition)**.** *For every $\mathcal{I}$-measurable $\boldsymbol{p}$,*

$$R(\boldsymbol{p} \mid \mathcal{I}) = d_K(\boldsymbol{p}, \boldsymbol{\mu}) + \frac{1}{K}\mathbb{E}[\|\boldsymbol{Y} - \boldsymbol{\mu}\|_2^2 \mid \mathcal{I}].$$

Proof. Expand

$$\|\boldsymbol{p} - \boldsymbol{Y}\|_2^2 = \|\boldsymbol{p} - \boldsymbol{\mu}\|_2^2 + \|\boldsymbol{Y} - \boldsymbol{\mu}\|_2^2 + 2\langle \boldsymbol{p} - \boldsymbol{\mu}, \boldsymbol{\mu} - \boldsymbol{Y}\rangle.$$

The conditional expectation of the cross term is zero because $\mathbb{E}[\boldsymbol{Y} - \boldsymbol{\mu} \mid \mathcal{I}] = 0$. Divide by K. □

Theorem A.10 (Correct-set conditional Brier guarantee)**.** *Let $\boldsymbol{p}^\star = P_{\mathcal{C}}(\boldsymbol{p}_0)$. If the conditional mean satisfies $\boldsymbol{\mu} \in \mathcal{C}$, then*

$$R(\boldsymbol{p}^\star \mid \mathcal{I}) \le R(\boldsymbol{p}_0 \mid \mathcal{I}) - d_K(\boldsymbol{p}_0, \boldsymbol{p}^\star). \tag{A.11}$$

Proof. Set $\boldsymbol{r} = \boldsymbol{p}_0 - \boldsymbol{p}^\star$. Since $\boldsymbol{\mu} \in \mathcal{C}$, the projection inequality gives

$$\langle \boldsymbol{r}, \boldsymbol{\mu} - \boldsymbol{p}^\star\rangle \le 0.$$

Expanding around $\boldsymbol{p}^\star$,

$$\|\boldsymbol{p}_0 - \boldsymbol{\mu}\|_2^2 = \|\boldsymbol{r}\|_2^2 + \|\boldsymbol{p}^\star - \boldsymbol{\mu}\|_2^2 + 2\langle \boldsymbol{r}, \boldsymbol{p}^\star - \boldsymbol{\mu}\rangle,$$

and the last term is nonnegative. Hence

$$\|\boldsymbol{p}^\star - \boldsymbol{\mu}\|_2^2 \le \|\boldsymbol{p}_0 - \boldsymbol{\mu}\|_2^2 - \|\boldsymbol{p}_0 - \boldsymbol{p}^\star\|_2^2.$$

Substitute this inequality into Lemma A.9 and divide by K. □

Theorem A.10 is conditional on the authority-feasible set containing the true conditional mean for the information set being used. It is not an unconditional statement that projection improves every observed label or every task metric.

We next quantify misspecification. Let

$$\epsilon = \operatorname{dist}(\boldsymbol{\mu}, \mathcal{C}) = \|\boldsymbol{\mu} - P_{\mathcal{C}}(\boldsymbol{\mu})\|_2.$$

Theorem A.11 (Cost of a misspecified authority set)**.** *For any nonempty closed convex $\mathcal{C}$,*

$$R(\boldsymbol{p}^\star \mid \mathcal{I}) - R(\boldsymbol{p}_0 \mid \mathcal{I}) \le \frac{\epsilon^2}{K}. \tag{A.12}$$

Proof. Let $\boldsymbol{z} = P_{\mathcal{C}}(\boldsymbol{\mu})$, $\boldsymbol{r} = \boldsymbol{p}_0 - \boldsymbol{p}^\star$, and $d = \|\boldsymbol{r}\|_2$. By Lemma A.9, the irreducible terms cancel, and

$$\begin{aligned} K[R(\boldsymbol{p}^\star \mid \mathcal{I}) - R(\boldsymbol{p}_0 \mid \mathcal{I})] &= \|\boldsymbol{p}^\star - \boldsymbol{\mu}\|_2^2 - \|\boldsymbol{p}_0 - \boldsymbol{\mu}\|_2^2 \\ &= -d^2 + 2\langle \boldsymbol{r}, \boldsymbol{\mu} - \boldsymbol{p}^\star\rangle. \end{aligned}$$

Decompose

$$\boldsymbol{\mu} - \boldsymbol{p}^\star = (\boldsymbol{\mu} - \boldsymbol{z}) + (\boldsymbol{z} - \boldsymbol{p}^\star).$$

Since $\boldsymbol{z} \in \mathcal{C}$, the projection inequality for $\boldsymbol{p}^\star = P_{\mathcal{C}}(\boldsymbol{p}_0)$ gives $\langle \boldsymbol{r}, \boldsymbol{z} - \boldsymbol{p}^\star\rangle \le 0$. Therefore

$$K\Delta R \le -d^2 + 2\langle \boldsymbol{r}, \boldsymbol{\mu} - \boldsymbol{z}\rangle \le -d^2 + 2d\epsilon \le \epsilon^2,$$

where the second inequality is Cauchy–Schwarz and the last follows from $-(d - \epsilon)^2 \le 0$. Dividing by K proves the result. □

The bound exposes rather than removes the cost of incorrect authority. For example, let $\boldsymbol{Y} = (1, 0)$, $\boldsymbol{p}_0 = (0.9, 0.1)$, and let an incorrect rule require $p_1 \le 0.2$ on the binary simplex. Then $\boldsymbol{p}^\star = (0.2, 0.8)$. The unnormalized squared error increases from 0.02 to 1.28, even though the rule is satisfied exactly. The normalized excess risk is 0.63, while $\epsilon^2/K = 1.28/2 = 0.64$, consistent with Theorem A.11.

A pointwise claim for an observed one-hot or binary label would require that particular label vector to belong to $\mathcal{C}$. The conditional result instead requires the conditional mean to belong to $\mathcal{C}$. These statements are not interchangeable. Neither theorem applies directly to mAP, R@1, MRR, or uncalibrated retrieval similarities.

A.7 ADMISSIBLE COMPUTATION

Decision feasibility alone cannot prevent an auxiliary source from changing a protected intermediate state. Fix the target, context, validity state, availability state, authority specification, and trained parameters. Partition the observed inputs into authorized inputs $\boldsymbol{x}_H$ and unauthorized auxiliary inputs $\boldsymbol{x}_L$. For the graph $\mathcal{G}_\mathcal{E}$ compiled by ASC, define the admissible protected-state class

$$\mathfrak{F}(\mathcal{G}_\mathcal{E}) = \{f : \boldsymbol{h}_H(\boldsymbol{x}_H, \boldsymbol{x}_L) = \boldsymbol{h}_H(\boldsymbol{x}_H, \boldsymbol{x}'_L) \ \ \forall \boldsymbol{x}_L, \boldsymbol{x}'_L\}\,. \tag{A.13}$$

This definition concerns the protected state, not necessarily the complete prediction. Auxiliary information may still affect unprotected states and the degrees of freedom left within the final authority-feasible set.

Theorem A.12 (APGM computational non-interference)**.** *Consider an L-layer message-passing network with fixed parameters. Suppose that:*

1. *every protected initial state depends only on authorized inputs;*
2. *every direct and indirect ancestor of a protected state belongs to the authorized subgraph or is an immutable support node;*
3. *attention normalization, gating, normalization statistics, caches, and readouts used to update a protected state depend only on those ancestors; and*
4. *immutable support nodes do not first absorb auxiliary sample information and then relay it into a protected state.*

Then, for any two forward passes that differ only in unauthorized auxiliary inputs,

$$\boldsymbol{h}_H^{(L)}(\boldsymbol{x}_H, \boldsymbol{x}_L) = \boldsymbol{h}_H^{(L)}(\boldsymbol{x}_H, \boldsymbol{x}'_L). \tag{A.14}$$

Consequently, the protected computation belongs to $\mathfrak{F}(\mathcal{G}_\mathcal{E})$.

Proof. At layer 0, protected states are equal because they depend only on the shared authorized inputs. Assume that all authorized ancestors of every protected state are equal at layer ℓ. By assumptions 2–4, the permitted neighbor set, attention logits, normalization terms, messages, gates, and pointwise updates used for the protected state are functions only of these equal ancestor states and fixed parameters. Its layer-$\ell + 1$ value is therefore equal in both passes. Induction over $\ell = 0, \ldots, L - 1$ proves equation A.14. □

Checking only whether a visible adjacency entry is zero is insufficient. Shared cross-source normalization, a global gate, a support-node relay, or a direct final-head bypass can violate the theorem while leaving the nominal protected edge masked. Stop-gradient also does not establish equation A.14: it changes backward derivatives but leaves forward information dependencies intact.

The guarantee is parameter-conditional. It does not claim that retraining on a different auxiliary distribution produces the same parameters. It also does not imply that the final prediction is invariant to auxiliary evidence. A final head may legitimately use auxiliary states, provided that ACP keeps the returned prediction inside the activated feasible set.

A.8 End-to-End Training and Prediction

Algorithm 1 End-to-End Training with Authority-Consistent Fusion

Require: Training data $\mathcal{S}_{\mathrm{tr}}$, validation data $\mathcal{S}_{\mathrm{val}}$, epochs T, projection weight $\lambda_{\mathrm{proj}} = 0.1$
Ensure: Validation-selected parameters $\theta^\star$

Initialize model parameters θ and optimizer
$\ell_{\mathrm{best}} \leftarrow +\infty$
for $t = 1, \dots, T$ **do**
  for minibatch $\mathcal{B} = \{(\mathcal{E}_n, \boldsymbol{y}_n)\}_{n=1}^{N}$ **do**
    $(\mathcal{G}_n, \boldsymbol{B}_n, \boldsymbol{b}_n, \sigma_n) \leftarrow \mathrm{ASC}(\mathcal{E}_n)$ for each n ▷ Label-free compilation
    $\boldsymbol{h}_n^0 \leftarrow \textsc{Encode}(\mathcal{E}_n; \theta)$
    for $\ell = 0, \dots, L-1$ **do**
      $\boldsymbol{h}_n^{\ell+1} \leftarrow \mathrm{APGM}(\boldsymbol{h}_n^{\ell}, \mathcal{G}_n; \theta)$
    end for
    $\boldsymbol{p}_{0,n} \leftarrow \textsc{Head}(\boldsymbol{h}_n^L, q_n, \mathcal{A}_{q_n}; \theta)$
    for each n with $\sigma_n \in \{\textsc{constrained}, \textsc{unconstrained}\}$ **do**
      if $\sigma_n = \textsc{constrained}$ **then**
        $\boldsymbol{p}_n^\star \leftarrow \arg\min_{\boldsymbol{p} \in \mathcal{D}_{q_n}} \frac{1}{2}\|\boldsymbol{p} - \boldsymbol{p}_{0,n}\|_2^2$
        subject to $\boldsymbol{B}_n \boldsymbol{p} \geq \boldsymbol{b}_n$
      else
        $\boldsymbol{p}_n^\star \leftarrow \boldsymbol{p}_{0,n}$
      end if
    end for
    Mark failed projections as $\sigma_n = \textsc{solver-failure}$
    $\mathcal{V} \leftarrow \{n : \sigma_n \in \{\textsc{constrained}, \textsc{unconstrained}\}\}$
    require $\mathcal{V} \neq \varnothing$
    $\mathcal{L}_{\mathrm{ACF}} \leftarrow \frac{1}{|\mathcal{V}|} \sum_{n \in \mathcal{V}} [d_K(\boldsymbol{p}_n^\star, \boldsymbol{y}_n) + \lambda_{\mathrm{proj}} d_K(\boldsymbol{p}_n^\star, \boldsymbol{p}_{0,n})]$
    Update θ by backpropagating $\nabla_\theta \mathcal{L}_{\mathrm{ACF}}$
  end for
  $\ell_{\mathrm{val}} \leftarrow \textsc{EvaluateBrier}(\theta, \mathcal{S}_{\mathrm{val}})$ ▷ Same compile–mix–project procedure
  if $\ell_{\mathrm{val}} < \ell_{\mathrm{best}}$ **then**
    $\theta^\star \leftarrow \theta$
    $\ell_{\mathrm{best}} \leftarrow \ell_{\mathrm{val}}$
  end if
end for
return $\theta^\star$

Prediction. At test time, the selected parameters $\theta^\star$ are fixed and the same compile–encode–mix–project procedure is applied once to each evidence record. A CONSTRAINED record returns the projected prediction $\boldsymbol{p}^\star$, while an UNCONSTRAINED record returns $\boldsymbol{p}^\star = \boldsymbol{p}_0$. Records with CONFLICT, NO-INPUT, or SOLVER-FAILURE status produce an abstention and are excluded from predictive-risk denominators. Each returned prediction is stored together with its status, activated relations, and source record, allowing structural compliance to be evaluated independently of predictive utility.

Algorithm 1 is end-to-end with respect to the learnable prediction path. Gradients from the projected Brier objective update the source encoders, APGM, and prediction head jointly through either the registered analytic projection or the differentiable general quadratic-program layer. ASC remains label-free and is not learned from the target labels. The algorithm makes no claim that the nonconvex outer training problem reaches a global optimum.

A.9 Unified Optimization View of ACF

The compiler produces

$$\mathrm{ASC}(\mathcal{E}) = (\mathcal{G}_{\mathcal{E}}, \mathcal{C}(\mathcal{E}), \sigma_{\mathcal{E}}),$$

where $\sigma_{\mathcal{E}}$ distinguishes constrained, unconstrained, conflict, no-input, and numerical-failure states. Let $\boldsymbol{p}_{0,\theta}(\mathcal{E})$ be the unconstrained prediction of a model whose protected computation satisfies equation A.13. The implemented learning problem can be written as

$$\min_{\theta:\, f_\theta \in \mathfrak{F}(\mathcal{G}_{\mathcal{E}}) \text{ a.s.}} \mathbb{E}\big[d_K\big(P_{\mathcal{C}(\mathcal{E})}(\boldsymbol{p}_{0,\theta}), \boldsymbol{Y}\big) + \lambda_{\mathrm{proj}} d_K\big(P_{\mathcal{C}(\mathcal{E})}(\boldsymbol{p}_{0,\theta}), \boldsymbol{p}_{0,\theta}\big)\big]. \tag{A.15}$$

The experiments use $\lambda_{\mathrm{proj}} = 0.1$. The first term learns predictive utility after enforcing the active decision relations. The second discourages systematic dependence on large corrections. The constraint on f_θ is implemented architecturally by APGM rather than by a soft penalty.

Equation A.15 separates two optimization levels. The inner decision problem is the convex projection in equation A.5 and has a unique solution whenever the compiled set is feasible. The outer neural-network problem is generally nonconvex. None of the projection results imply that stochastic training finds a global optimum of the outer problem.

Theorem A.13 (ACF composition)**.** *Fix an evidence record for which:*

1. *ASC is label-free and returns a nonempty closed convex set $\mathcal{C}(\mathcal{E})$;*
2. *the APGM ancestor-closure conditions of Theorem A.12 hold; and*
3. *ACP exactly solves equation A.5.*

Then ACF satisfies all of the following:

1. ***Computational non-interference:*** *the protected state is invariant to interventions that change only unauthorized auxiliary inputs while holding the compilation fixed;*
2. ***Decision feasibility:*** $\boldsymbol{p}^\star \in \mathcal{C}(\mathcal{E})$*;*
3. ***Unique minimum intervention:*** *no other authority-feasible prediction is closer to $\boldsymbol{p}_0$ in Euclidean distance;*
4. ***Fixed-set stability:*** *the ACP mapping is firmly non-expansive and hence non-expansive; and*
5. ***Conditional risk control:*** *if $\mathbb{E}[\boldsymbol{Y} \mid \mathcal{I}] \in \mathcal{C}(\mathcal{E})$, then equation A.11 holds, while a misspecified set obeys equation A.12.*

Proof. Statement 1 is Theorem A.12. Statements 2 and 3 follow from the definition of equation A.5 and Theorem A.3. Statement 4 is Theorem A.7. Statement 5 follows from Theorems A.10 and A.11. □

The theorem is modular: failure of one premise removes only the corresponding guarantee. If $\mathcal{C} = \varnothing$, ACP is undefined and ACF abstains. If the compiler is feasible but a numerical solver fails, no feasibility guarantee is claimed. If APGM admits an unauthorized bypass, the projection may still enforce the final decision relation, but protected-state non-interference is lost. If the authority specification is incorrect, the model may remain perfectly compliant while incurring the risk quantified in Theorem A.11.

Finally, the result is limited to affine relations defining nonempty closed convex feasible sets. Disjunctive, combinatorial, or otherwise nonconvex authority rules may produce multiple nearest feasible predictions and need not yield a non-expansive decision operator. Learned modality weights or graph edges also do not become authority relations merely because they are stable or predictive. Authority remains an externally specified component of the task, while optimization determines how that specification constrains computation and decisions.

B Complete Experiments, Interventions, and Evidence Audit

This appendix expands the experiments around the three claims made in the main text rather than the chronological order in which the systems were developed. Sections B.1–B.3 test whether authority can be specified independently of predictive utility and reliability. Sections B.4–B.5 audit admissible computation and decisions. Section B.6 separates representation geometry, communication, and edge allocation on UTSW MRI-to-semantic retrieval. The controlled tasks are mechanism studies, whereas UTSW is a structural attribution setting; neither is presented as patient-level validation of clinical authority.

Table B.1: Authority Identification Suite and Evaluation Protocol

Task	Output domain	Classes	Relations	Train/Val/Test	Events	Data seeds	Model seeds
S1	Simplex	4	1	12,000/3,000/6,000	42	5	3
S2	Hypercube	6	2	12,000/3,000/6,000	14	5	3

The formal study comprises 300 main and 210 attribution training runs. Ten rule-only evaluations do not train a model. A separate 60-run development search selected learning rates and is excluded from all reported uncertainty summaries.

Table B.2: Information-matched baselines: which variable does each method encode?

Method	Authority information	Communication	Protected mask	Reliability/dominance	Projection	Objective
Rule-only	Policy, states, statements	None	N/A	None	Fixed	N/A
Uniform	All	Fixed uniform	No	None	No	Brier
Learned	All	Learned graph	No	None	No	Brier
Reliability	All	Quality-weighted	No	Reliability	No	Brier
Dominant	All	Dominant source	No	Dominance	No	Brier
Transformer	All	Self-attention	No	Learned	No	Brier
Graph-only	All	Allowed graph	Yes	None	No	Brier
Soft	All	Learned graph	No	None	Soft penalty	Brier + relation
Learned-joint	All	Learned graph	No	None	Joint	Projected Brier
Transformer-joint	All	Self-attention	No	Learned	Joint	Projected Brier
ACF	All	Allowed graph	Yes	Separate	Joint	Projected Brier

All learned baselines receive the same policy, validity, availability, quality, and structured statements. Post-hoc variants reuse the corresponding unprojected checkpoint and therefore are not additional trained methods. "Joint" means projection is present during training; it does not imply protected computation.

B.1 Controlled Tasks and Evaluation Protocol

Tasks. S1 is a four-class problem on the simplex. A task-conditioned policy activates one admissible class relation for each evidence record. S2 is a six-label problem on the hypercube with two relation-specific probability-order constraints. The authorized source varies by record in both tasks. Every strong baseline receives the same policy, source validity, availability, quality, and structured statements; the comparison therefore does not reward ACF for privileged information. Policy compilation never reads the current label.

Replications and selection. Each formal configuration uses five independently generated datasets and three model initializations per dataset. We first average model seeds within a data seed and treat the resulting five values as the statistical units. Learning-rate development was conducted separately and is not pooled with the 510 formal training runs. UTSW uses three predefined splits and three initializations; its intervals describe variation under that protocol, because the splits originate from the same cohort and share some test patients.

Metrics. For K outputs, Brier is $K^{-1}\|\boldsymbol{p}-\boldsymbol{y}\|_2^2$. Residual Brier and residual accuracy evaluate degrees of freedom not fixed by an activated relation. Authority-violation rate (AVR) divides violated activated relations by the number evaluable; an empty denominator is N/A. Coverage is the fraction of records returning a prediction. Protected-state shift is the paired maximum or mean change in a protected representation after an auxiliary intervention. Projection MSE is $K^{-1}\|\boldsymbol{p}^\star-\boldsymbol{p}_0\|_2^2$. Conflict, no-input, and solver-failure records are excluded from predictive-risk denominators but retained in status and coverage audits.

B.2 Complete Results on the Authority Identification Suite

Table B.3 reports every formal clean baseline. The essential comparison is not whether a constrained method always minimizes Brier. It is whether predictive utility is retained while authority compliance is made explicit. On S1, unprojected graph, uniform, learned, reliability, dominance, transformer, and soft baselines violate the active relation for every evaluable record, although several have competitive Brier. Projection removes those decision violations. On S2 the clean event is non-diagnostic for AVR because the learned outputs already satisfy the activated orders; later interventions provide the necessary stress tests.

Table B.3: Complete clean results: predictive utility and structural compliance are reported separately.

Method	Brier ↓	Resid. Brier ↓	Acc. ↑	Resid. Acc. ↑	AVR ↓	Coverage	Proj. MSE
Panel 1. Task S1							
Transformer-joint	**0.080724 ± 0.001313**	0.161448 ± 0.002626	0.801689 ± 0.004246	0.801689 ± 0.004246	0.000 ± 0.000	1.000 ± 0.000	0.001572 ± 0.001297
Learned-joint	0.082434 ± 0.001608	0.164869 ± 0.003217	0.798256 ± 0.004839	0.798256 ± 0.004839	0.000 ± 0.000	1.000 ± 0.000	0.001740 ± 0.000600
ACF	0.082380 ± 0.001363	0.164760 ± 0.002727	0.798356 ± 0.004046	0.798356 ± 0.004046	0.000 ± 0.000	1.000 ± 0.000	0.001651 ± 0.000763
Graph-only	0.084086 ± 0.001489	0.168236 ± 0.002998	0.794811 ± 0.004694	0.794767 ± 0.004664	1.000 ± 0.000	1.000 ± 0.000	0.000029 ± 0.000006
Uniform graph	0.082827 ± 0.001324	0.165651 ± 0.002661	0.799322 ± 0.003700	0.799344 ± 0.003687	1.000 ± 0.000	1.000 ± 0.000	0.000044 ± 0.000008
Learned graph	0.083114 ± 0.001502	0.166278 ± 0.003007	0.797600 ± 0.005215	0.797633 ± 0.005229	1.000 ± 0.000	1.000 ± 0.000	0.000048 ± 0.000009
Reliability-weighted	0.082737 ± 0.001709	0.165523 ± 0.003438	0.799067 ± 0.004608	0.799078 ± 0.004628	1.000 ± 0.000	1.000 ± 0.000	0.000040 ± 0.000005
Dominant-source	0.083125 ± 0.001286	0.166301 ± 0.002616	0.796644 ± 0.003300	0.796656 ± 0.003301	1.000 ± 0.000	1.000 ± 0.000	0.000041 ± 0.000006
Transformer	0.081816 ± 0.001314	0.163498 ± 0.002688	0.800322 ± 0.003695	0.800344 ± 0.003666	1.000 ± 0.000	1.000 ± 0.000	0.000124 ± 0.000108
Soft constraint	0.083201 ± 0.001471	0.166451 ± 0.002947	0.797333 ± 0.004631	0.797344 ± 0.004639	1.000 ± 0.000	1.000 ± 0.000	0.000040 ± 0.000004
Rule-only	0.125000 ± 0.000000	0.250000 ± 0.000000	0.501167 ± 0.006885	0.498833 ± 0.006885	0.000 ± 0.000	1.000 ± 0.000	0.062500 ± 0.000000
Graph-only + post-hoc	0.084057 ± 0.001485	0.168114 ± 0.002970	0.794811 ± 0.004694	0.794811 ± 0.004694	0.000 ± 0.000	1.000 ± 0.000	0.000029 ± 0.000006
Learned + post-hoc	0.083065 ± 0.001500	0.166131 ± 0.003001	0.797600 ± 0.005215	0.797600 ± 0.005215	0.000 ± 0.000	1.000 ± 0.000	0.000048 ± 0.000009
Transformer + post-hoc	0.081692 ± 0.001317	0.163383 ± 0.002633	0.800322 ± 0.003695	0.800322 ± 0.003695	0.000 ± 0.000	1.000 ± 0.000	0.000124 ± 0.000108
Panel 2. Task S2							
Transformer-joint	0.197479 ± 0.001003	0.217515 ± 0.001120	0.719591 ± 0.001945	0.659139 ± 0.002559	0.000 ± 0.000	1.000 ± 0.000	0.000000 ± 0.000000
Learned-joint	0.196741 ± 0.001192	0.215642 ± 0.001510	0.721239 ± 0.002443	0.664050 ± 0.003710	0.000 ± 0.000	1.000 ± 0.000	0.000000 ± 0.000000
ACF	0.196675 ± 0.001135	0.215444 ± 0.001374	0.721231 ± 0.002584	0.664033 ± 0.004255	0.000 ± 0.000	1.000 ± 0.000	0.000000 ± 0.000000
Graph-only	0.196835 ± 0.001090	0.215657 ± 0.001300	0.721370 ± 0.002458	0.664450 ± 0.003933	0.000 ± 0.000	1.000 ± 0.000	0.000000 ± 0.000000
Uniform graph	**0.196438 ± 0.001063**	0.214548 ± 0.001233	0.721956 ± 0.002188	0.666206 ± 0.002936	0.000 ± 0.000	1.000 ± 0.000	0.000000 ± 0.000000
Learned graph	0.196934 ± 0.001115	0.215965 ± 0.001327	0.721006 ± 0.002393	0.663356 ± 0.003941	0.000 ± 0.000	1.000 ± 0.000	0.000000 ± 0.000000
Reliability-weighted	0.196838 ± 0.001111	0.215714 ± 0.001362	0.721163 ± 0.002443	0.663822 ± 0.004101	0.000 ± 0.000	1.000 ± 0.000	0.000000 ± 0.000000
Dominant-source	0.196875 ± 0.001161	0.215697 ± 0.001459	0.721224 ± 0.002450	0.664011 ± 0.004041	0.000 ± 0.000	1.000 ± 0.000	0.000000 ± 0.000000
Transformer	0.197567 ± 0.000973	0.217750 ± 0.001070	0.719613 ± 0.001889	0.659211 ± 0.002046	0.000 ± 0.000	1.000 ± 0.000	0.000000 ± 0.000000
Soft constraint	0.196942 ± 0.001112	0.215959 ± 0.001333	0.721070 ± 0.002394	0.663550 ± 0.004018	0.000 ± 0.000	1.000 ± 0.000	0.000000 ± 0.000000
Rule-only	0.250000 ± 0.000000	0.250000 ± 0.000000	0.502044 ± 0.002143	0.501467 ± 0.005546	0.000 ± 0.000	1.000 ± 0.000	0.000000 ± 0.000000
Graph-only + post-hoc	0.196835 ± 0.001090	0.215657 ± 0.001300	0.721370 ± 0.002458	0.664450 ± 0.003933	0.000 ± 0.000	1.000 ± 0.000	0.000000 ± 0.000000
Learned + post-hoc	0.196934 ± 0.001115	0.215965 ± 0.001327	0.721006 ± 0.002393	0.663356 ± 0.003941	0.000 ± 0.000	1.000 ± 0.000	0.000000 ± 0.000000
Transformer + post-hoc	0.197567 ± 0.000973	0.217750 ± 0.001070	0.719613 ± 0.001889	0.659211 ± 0.002046	0.000 ± 0.000	1.000 ± 0.000	0.000000 ± 0.000000

Entries are mean ± sample SD over five data-seed means; each learned data-seed mean averages three model seeds. Indented post-hoc rows reuse the same checkpoint. Only the lowest Brier per task is bold. S2 clean AVR is zero for every method and has no discriminative value in that condition. Coverage, AVR, and Brier answer different questions and are not jointly ranked.

Table B.4 uses matched data seeds rather than treating all model runs as independent observations. ACF and learned-joint differ by at most a small fraction of the Brier scale, which is evidence of matched utility, not general superiority. Comparisons with graph-only isolate projection from protected computation, whereas comparisons with learned and uniform models show why predictive performance alone cannot reveal the authority mechanism.

Table B.4: Paired predictive comparisons: does ACF preserve utility under matched authority information?

Comparator	Mean ΔBrier	95% bootstrap CI	Data-seed direction	Interpretation
Task S1: ACF minus comparator				
Learned-joint	-0.000055	[-0.000563, +0.000537]	2+/3-	Matched utility; not an architecture win
Transformer-joint	+0.001656	[+0.001164, +0.002147]	5+/0-	Predictive comparison under matched information
Graph-only	-0.001706	[-0.002458, -0.000953]	0+/5-	Adds decision projection to protected graph
Learned graph	-0.000734	[-0.001114, -0.000171]	1+/4-	Predictive comparison under matched information
Uniform graph	-0.000447	[-0.000896, +0.000002]	1+/4-	Predictive comparison under matched information
Task S2: ACF minus comparator				
Learned-joint	-0.000066	[-0.000126, -0.000006]	1+/4-	Matched utility; not an architecture win
Transformer-joint	-0.000804	[-0.000985, -0.000634]	0+/5-	Predictive comparison under matched information
Graph-only	-0.000160	[-0.000250, -0.000079]	0+/5-	Adds decision projection to protected graph
Learned graph	-0.000259	[-0.000326, -0.000192]	0+/5-	Predictive comparison under matched information
Uniform graph	+0.000237	[+0.000144, +0.000331]	5+/0-	Predictive comparison under matched information

Negative differences favor ACF. Differences are formed after averaging the three model seeds within each of five data seeds; percentile bootstrap intervals resample those five independent generated datasets (20,000 draws, seed 20260921). No significance stars are used.

B.3 Authority Interventions and Failure States

Authority is externally specified and task-conditioned, so it should change under policy reassignment even when predictive strength or measured reliability is held fixed. Conversely, a reliability intervention may change predictive risk without redefining which relation is entitled to constrain the decision. Table B.5 evaluates incorrect, shuffled, and reassigned specifications while preserving this distinction. Wrong-policy and wrong-statement conditions keep the ground-truth task fixed and expose the cost of activating invalid relations. Target switch deliberately changes the task and therefore cannot be read as an ordinary robustness perturbation.

Figure 4 complements the aggregate tables with the five independent data-seed units. The paired clean trajectories show that ACF and learned-joint retain closely matched predictive utility, whereas the intervention panel shows the risk incurred by an incorrect specification. The figure does not treat the three model initializations within a data seed as independent datasets.

Table B.5: Specification interventions: what changes when authority is incorrect or reassigned?

Condition	Brier	ΔBrier vs. clean	AVR	Coverage	Protected shift
Task S1					
Clean	0.082380	+0.000000	0.000	1.000	0.000000
Wrong policy	0.271208	+0.188828	0.000	1.000	N/A
Wrong accepted statement	0.421521	+0.339141	0.000	1.000	N/A
Shuffled policy	0.224043	+0.141663	0.000	1.000	0.000000
Target switch	0.082369	task redefinition	0.000	1.000	N/A
Task S2					
Clean	0.196675	+0.000000	0.000	1.000	0.000000
Wrong policy	0.280580	+0.083905	0.000	1.000	N/A
Wrong accepted statement	0.364774	+0.168098	0.000	1.000	N/A
Shuffled policy	0.259447	+0.062772	0.000	1.000	0.000000
Target switch	0.196405	task redefinition	0.000	1.000	N/A

Means are computed from five data-seed means. Wrong-policy, wrong-statement, and shuffled-policy evaluations preserve the true task labels, so their deltas measure misspecification cost. Target switch changes both target definition and labels and is therefore reported as task redefinition, not fixed-label robustness.

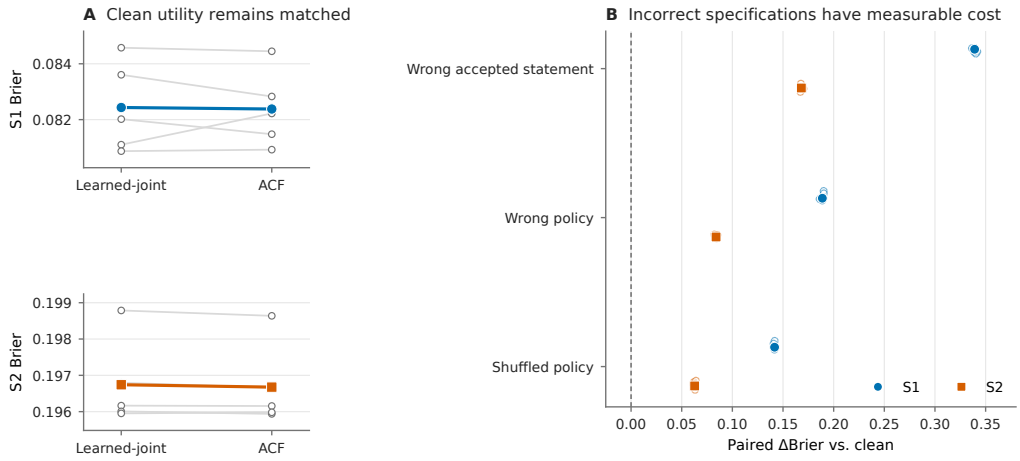


Figure 4: Matched clean utility and authority misspecification cost. Panel A connects the learned-joint and ACF Brier scores for each of five independently generated datasets; colored segments connect the corresponding means. Panel B reports paired event-minus-clean differences, showing individual data-seed values and 95% bootstrap intervals. Lower Brier is better; positive intervention deltas indicate higher risk.

Table B.6 audits the compiler's state machine. Invalid or missing evidence withdraws its hard constraint instead of filling a zero row into the relation matrix. Unknown evidence can still enter an auxiliary path but cannot authorize a hard relation. Infeasible relation systems and absent inputs produce abstention. This separates a mathematical conflict, a deliberate no-input decision, and a numerical solver failure.

The factor study in Table B.7 and Figure 5 changes statement error, auxiliary reliability, and conflict strength without using them to infer authority. The response curves reveal that reliability and statement error dominate the observed Brier variation, while the nearly overlapping κ curves show that conflict weight contributes little over the tested range.

B.4 Computational Compliance and Component Attribution

Decision projection cannot by itself establish computational non-interference. We therefore perturb auxiliary evidence while matching the target, authorized source, relation direction, and activation state. A valid APGM leaves the protected state unchanged. As a positive control, a shared gate is

Table B.6: Validity, availability, and failure states: when is a prediction returned?

Condition	Active rel./record	Coverage	Returned Brier	AVR denom.	AVR	Pooled status count
Task S1						
Clean	1.000	1.000	0.082380	90000	0.000	constrained=90000
Invalid	0.000	1.000	0.199872	0	N/A	unconstrained=90000
Unknown	0.000	1.000	0.186821	0	N/A	unconstrained=90000
Missing	0.000	1.000	0.200192	0	N/A	unconstrained=90000
All missing	0.000	0.000	N/A	0	N/A	no-input=90000
Conflict	3.000	0.000	N/A	0	N/A	conflict=90000
Solver failure	N/A	N/A	N/A	0	N/A	0
Task S2						
Clean	2.000	1.000	0.196675	180000	0.000	constrained=90000
Invalid	0.000	1.000	0.249232	0	N/A	unconstrained=90000
Unknown	0.000	1.000	0.247851	0	N/A	unconstrained=90000
Missing	0.000	1.000	0.248841	0	N/A	unconstrained=90000
All missing	0.000	0.000	N/A	0	N/A	no-input=90000
Conflict	4.000	0.000	N/A	0	N/A	conflict=90000
Solver failure	N/A	N/A	N/A	0	N/A	0

Counts pool the 15 ACF runs per task; "active rel./record" is a count and can exceed one under deliberately contradictory conflict events. Returned Brier is averaged over run-level values and excludes abstentions. Invalid and missing evidence withdraw the corresponding hard relation; unknown evidence can remain auxiliary but does not activate it. Conflict and all-missing states have zero coverage. An empty AVR denominator is N/A, not zero. Solver failures are counted from saved run records and are distinct from mathematical conflict.

Table B.7: Factor interventions: reliability changes utility without defining authority.

Panel A. S1 factorial Brier

η	κ	$r_b = .60$	$r_b = .80$	$r_b = .95$
0.00	0.0	0.140672	0.082310	0.038067
0.00	0.5	0.140612	0.082481	0.038447
0.00	1.0	0.140648	0.082521	0.038436
0.05	0.0	0.153908	0.098514	0.056302
0.05	0.5	0.153849	0.098695	0.056696
0.05	1.0	0.153853	0.098717	0.056672
0.15	0.0	0.181499	0.132108	0.094203
0.15	0.5	0.181452	0.132278	0.094561
0.15	1.0	0.181478	0.132319	0.094575

Panel B. S2 reliability

r_b	Brier	AVR	Shift
0.60	0.202673	0.000	0.000000
0.80	0.196675	0.000	0.000000
0.95	0.196919	0.000	0.000000

Panel C. S1 conflict strength

Strength	Brier	Resid.	AVR
1	0.082521	0.165041	0.000
2	0.084814	0.169628	0.000
4	0.108106	0.216212	0.000
8	0.154426	0.308853	0.000

Cells average five data-seed means after model-seed averaging. In Panel A, η is statement-error probability, r_b auxiliary reliability, and κ auxiliary-conflict weight. These factors diagnose utility and conflict sensitivity while the authority policy remains fixed.

deliberately inserted to create a bypass; Table B.8 shows that the audit detects the resulting leakage. The positive control is not a new baseline and its predictive score is secondary to its known structural violation.

Figure 6 expands the compliance audit beyond its mean values. The protected-state panel displays all trained runs lightly and the five data-seed means prominently; these marks are not interpreted as 15 independent datasets. The coverage panel verifies that mathematical conflict and no-input states abstain rather than being silently scored as predictions.

Table B.9 separates implementation choices from the authority principle. The Euclidean encoder is not worse than the SPD encoder in either controlled task; hence SPD geometry cannot explain authority compliance. Removing anchors and shifting role masks have task-dependent effects. The low-sample comparison is retained because it was prespecified, but any advantage is restricted to the task and directions actually observed.

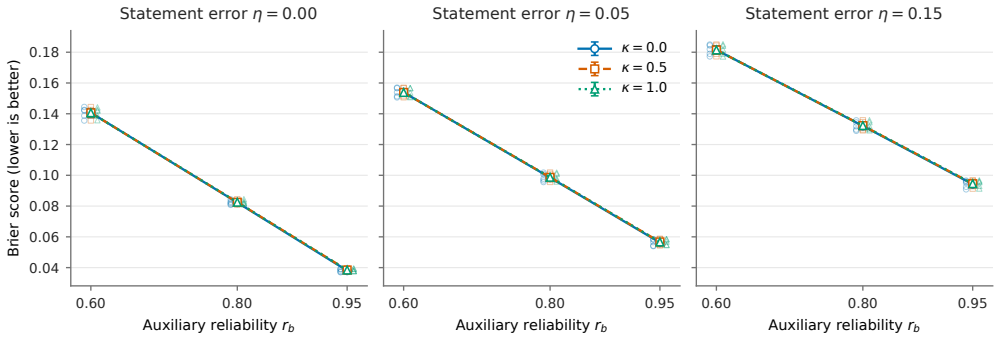


Figure 5: S1 factorial response curves across statement error η, auxiliary reliability r_b, and conflict weight κ. Open points are the five data-seed means after model-seed averaging; lines and error bars show their mean and 95% bootstrap interval. Marker shape and line style duplicate the color encoding. Lower Brier is better.

Table B.8: Protected-state audit: can the metric detect an unauthorized computational path?

Task	Condition	Shift mean	Shift max	Brier	AVR
S1	Valid APGM	0.000000	0.000000	0.082380	0.000
S1	Deliberate bypass	0.062627	0.171089	0.156292	0.000
S2	Valid APGM	0.000000	0.000000	0.196675	0.000
S2	Deliberate bypass	0.056688	0.124748	0.233645	0.000

The positive control deliberately introduces a shared gate into the same network. Shifts are computed only for pairs with the same target, authorized source, relation direction, and activation state. The bypass is an audit probe, not a competing method.

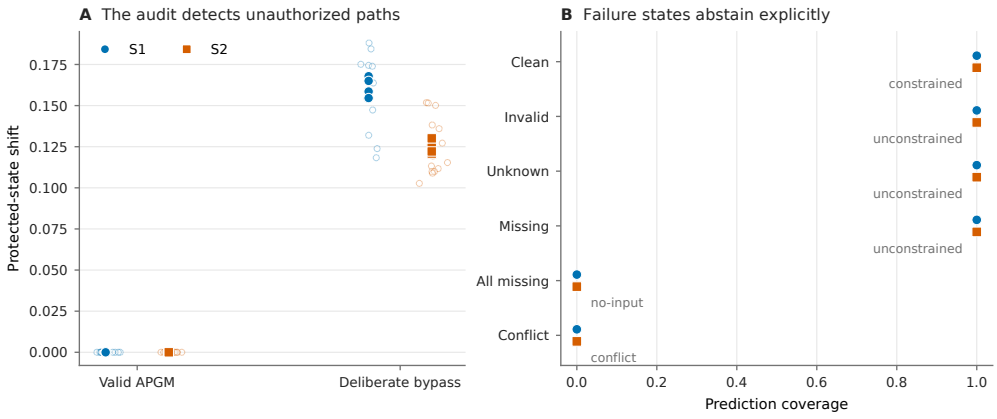


Figure 6: Computational compliance and explicit failure handling. Panel A contrasts the zero protected-state shift under valid APGM computation with the deliberate shared-gate bypass positive control. Small open marks are model runs and filled marks are data-seed means. Panel B reports prediction coverage under validity, availability, no-input, and conflict events; the adjacent labels name the compiled status rather than a predictive outcome.

Table B.9: Representation and architecture ablations: which implementation choices are task-dependent?

Contrast	Mean ΔBrier	95% bootstrap CI	Direction	Interpretation
Task S1				
ACF Euclidean − ACF SPD	-0.001012	[-0.001384, -0.000543]	0+/5-	Geometry is not authority
Learned-joint Euclidean − SPD	-0.001528	[-0.001979, -0.001077]	0+/5-	Representation control
ACF no anchors − ACF	+0.001034	[+0.000547, +0.001509]	5+/0-	Support-node contribution
Learned-joint no anchors − full	-0.000410	[-0.000771, -0.000015]	1+/4-	Architecture-matched anchor control
ACF role shift − ACF	+0.001419	[+0.000982, +0.001858]	5+/0-	Role-mask sensitivity
ACF low sample − learned-joint low sample	-0.004823	[-0.009847, -0.001307]	0+/5-	Low-sample comparison
Task S2				
ACF Euclidean − ACF SPD	-0.000688	[-0.000822, -0.000481]	0+/5-	Geometry is not authority
Learned-joint Euclidean − SPD	-0.000760	[-0.000918, -0.000561]	0+/5-	Representation control
ACF no anchors − ACF	+0.000656	[+0.000550, +0.000767]	5+/0-	Support-node contribution
Learned-joint no anchors − full	-0.000016	[-0.000110, +0.000077]	3+/2-	Architecture-matched anchor control
ACF role shift − ACF	+0.000070	[+0.000009, +0.000121]	4+/1-	Role-mask sensitivity
ACF low sample − learned-joint low sample	+0.000141	[-0.000078, +0.000277]	4+/1-	Low-sample comparison

Differences use five paired data-seed means. Euclidean representations are no worse than SPD in these controlled tasks, so SPD is not an authority contribution. Anchor and role-mask effects vary by task. The low-sample advantage is interpreted only where its data-seed direction is consistent; it is not a universal sample-efficiency claim.

B.5 Efficiency and Optimization Behavior

Table B.10 reports only quantities recorded in the controlled runs. Training time, inference latency, parameter count, selected validation epoch, and projection displacement are descriptive; they do not establish compute-matched superiority. We do not estimate missing FLOPs. This table also prevents the structural compliance result from being conflated with an unmeasured efficiency claim.

B.6 UTSW MRI-to-Semantic Retrieval

UTSW is used to attribute predictive differences among representation geometry, communication budget, and learned edge allocation. The outcome is MRI-to-semantic retrieval mAP; R@1 and MRR are secondary query-level metrics. Eight confirmation variants use all three splits and all three initializations. HeMIS, GMU, and MBT-style were used only in the prespecified screening with one initialization per split, so their $n = 3$ rows are explicitly separated from the $n = 9$ confirmation evidence.

The paired contrasts in Table B.12 match split and initialization. SPD improves over its Euclidean counterpart, fixed uniform communication improves over identity, and learned allocation underperforms the matched uniform budget. These effects are controlled contrasts under the present pipeline, not additive causal components or universal claims about geometric learning. The learned checkpoint also degrades when its graph is replaced at test time, while a model retrained with that replacement performs better. This direction reversal is the co-adaptation diagnostic: intervention sensitivity alone cannot identify a recovered topology, much less authority.

Figure 7 makes the evidential hierarchy explicit. Confirmation-model scores retain their nine protocol repetitions, while screening-only baselines are omitted from the ranking panel. The effect panel uses matched split and initialization pairs and keeps retrained contrasts separate from learned-checkpoint interventions.

Subgroup and stability audits further limit the interpretation. Support-node gains concentrate in molecular retrieval rather than appearing uniformly across target families. Edge correlations vary by split and layer, and the upper graph is partly structured by fixed support blocks. Even stable learned edges would show reproducibility under this parameterization, not correspondence to an unobserved graph truth.

The direction reversal is expanded in Figure 8. Replacing a learned graph at test time reduces mAP, but training and testing under the replacement produces the strongest of the three displayed conditions. The subgroup and edge panels then delimit the interpretation: support-node effects depend on the target family, and edge reproducibility varies across splits and layers.

B.7 Completeness and Reproducibility Audit

All formal authority jobs and all scheduled UTSW keys are present. Validation-only selection is preserved, and the saved UTSW scores were independently recomputed from patient-level direct scores and targets. Historical discovery directories for geodesic, hierarchical-SPD, and graph-evidence studies are provenance sources only; they do not define appendix sections or add statistical units. Table B.14 records the exact completion and audit scope.

Scope of the evidence. The controlled experiments identify the behavior of an externally specified authority mechanism under known data-generating rules. UTSW supplies a high-risk, structured retrieval setting for representation and graph attribution. Neither establishes that a learned topology is clinically true or that an authority policy is clinically valid. Those are external-specification and deployment questions beyond the guarantees and experiments reported here.

Table B.10: Computational characteristics on the Authority Identification Suite: what is the measured optimization cost?

Method	Parameters	Train sec.	Inference ms/sample	Best val. epoch	Projection MSE
Task S1					
Uniform graph	46,322	56.8	0.048	43.0	0.000044
Learned graph	46,322	66.0	0.049	22.0	0.000048
Reliability-weighted	46,322	67.4	0.051	19.7	0.000040
Dominant-source	46,322	66.7	0.050	25.1	0.000041
Transformer	129,970	69.3	0.049	6.7	0.000124
Graph-only	46,322	68.7	0.049	46.6	0.000029
Soft constraint	46,322	77.6	0.054	19.7	0.000040
Learned-joint	46,322	72.3	0.052	18.0	0.001740
Transformer-joint	129,970	74.4	0.049	6.1	0.001572
ACF	46,322	71.5	0.048	19.9	0.001651
Rule-only	N/A	N/A	0.039	N/A	0.062500
Task S2					
Uniform graph	46,722	59.6	0.055	41.9	0.000000
Learned graph	46,722	71.5	0.060	42.6	0.000000
Reliability-weighted	46,722	70.5	0.057	42.6	0.000000
Dominant-source	46,722	72.0	0.057	42.2	0.000000
Transformer	130,370	76.7	0.056	28.8	0.000000
Graph-only	46,722	72.3	0.054	41.1	0.000000
Soft constraint	46,722	78.8	0.057	41.7	0.000000
Learned-joint	46,722	78.6	0.059	39.3	0.000000
Transformer-joint	130,370	81.2	0.056	26.3	0.000000
ACF	46,722	77.0	0.054	38.8	0.000000
Rule-only	N/A	N/A	0.028	N/A	0.000000

Descriptive means use the same server environment and 15 runs per learned method and task. Rule-only has no training phase. FLOPs were not recorded and are not imputed. Equal training schedules do not imply parameter- or compute-matched architectures.

Table B.11: UTSW MRI-to-semantic retrieval: confirmation and screening have different evidential strength.

Panel A. Confirmation models: three splits $\times$ three model seeds

Variant	n	Validation mAP	Test mAP	R@1	MRR
spd_ub	9	0.7110 ± 0.0092	0.7267 ± 0.0094	0.7605 ± 0.0148	0.8469 ± 0.0106
spd_uniform_graph	9	0.7118 ± 0.0113	0.7256 ± 0.0104	0.7575 ± 0.0120	0.8447 ± 0.0100
spd_upper_only	9	0.7002 ± 0.0088	0.7127 ± 0.0105	0.7386 ± 0.0166	0.8345 ± 0.0136
spd_cross_graph	9	0.6911 ± 0.0120	0.7054 ± 0.0140	0.7527 ± 0.0210	0.8462 ± 0.0157
spd_identity_graph	9	0.6855 ± 0.0101	0.7007 ± 0.0097	0.7506 ± 0.0165	0.8455 ± 0.0106
spd_no_anchor_family	9	0.6739 ± 0.0133	0.6867 ± 0.0131	0.7413 ± 0.0242	0.8392 ± 0.0166
euclidean_cross_graph	9	0.6241 ± 0.0164	0.6405 ± 0.0202	0.7413 ± 0.0145	0.8429 ± 0.0086
latent_concat	9	0.6150 ± 0.0088	0.6298 ± 0.0108	0.7171 ± 0.0289	0.8313 ± 0.0185

Panel B. Screening-only baselines: split seeds 42/43/44, model seed 101

Variant	n	Validation mAP	Test mAP
latent_concat	3	0.6112 ± 0.0023	0.6256 ± 0.0025
mbt_style	3	0.6105 ± 0.0132	0.6259 ± 0.0125
hemis	3	0.6081 ± 0.0056	0.6206 ± 0.0062
gmu	3	0.6078 ± 0.0089	0.6167 ± 0.0125

Entries are mean $\pm$ sample SD over protocol runs. Screening-only HeMIS, GMU, and MBT-style results ($n = 3$) are not presented as evidence of equal strength to the $n = 9$ confirmation models. Model selection used validation only. UTSW diagnoses representation and graph attribution; it is not patient-level validation of clinical authority.

Table B.12: Matched UTSW contrasts: can geometry, communication, and allocation be separated?

Panel A. Retrained paired contrasts

Contrast	ΔmAP points	95% CI	Positive
Geometry: SPD learned - Euclidean learned	+6.485	[+4.588, +9.055]	9/9
Communication: uniform-both - identity-both	+2.482	[+1.812, +3.148]	9/9
Allocation: learned-both - uniform-both	-2.018	[-2.890, -1.369]	0/9
SPD-UB - learned-both	+2.128	[+1.437, +3.084]	9/9
SPD-UB - latent concat	+9.687	[+9.065, +10.300]	9/9
SPD-UB - uniform-both	+0.111	[-0.076, +0.289]	7/9
Upper communication: SPD-UB - identity-both	+2.593	[+1.951, +3.242]	9/9
Local learned: learned-both - upper-only	-0.735	[-1.326, -0.079]	2/9
Upper allocation: upper-only - SPD-UB	-1.394	[-2.277, -0.555]	1/9
Anchor family: full - no-anchor-family	+1.867	[+1.208, +2.399]	9/9

Panel B. Learned-checkpoint interventions

Graph	mAP	Δ points	95% CI	Drops
identity	0.6919	-1.346	[-2.168, -0.658]	9/9
no_local	0.6974	-0.794	[-1.222, -0.348]	7/9
no_region_family	0.7032	-0.216	[-0.707, +0.170]	4/9
no_upper	0.6993	-0.613	[-1.325, -0.130]	8/9
shuffle	0.6930	-1.243	[-2.565, -0.257]	8/9
uniform	0.6706	-3.474	[-7.209, -0.667]	8/9

Contrasts match split and model seed. For A = learned-both and B = uniform-both, $S(A, A) = 0.705391$, $S(A, B) = 0.670648$, and $S(B, B) = 0.725569$. Thus intervention cost is $+3.474$ points, retrained advantage is $+2.018$ points, and the co-adaptation gap is $+5.492$ points. Test-time sensitivity therefore does not establish topology recovery. Intervals describe current protocol variation, not a patient-population CI.

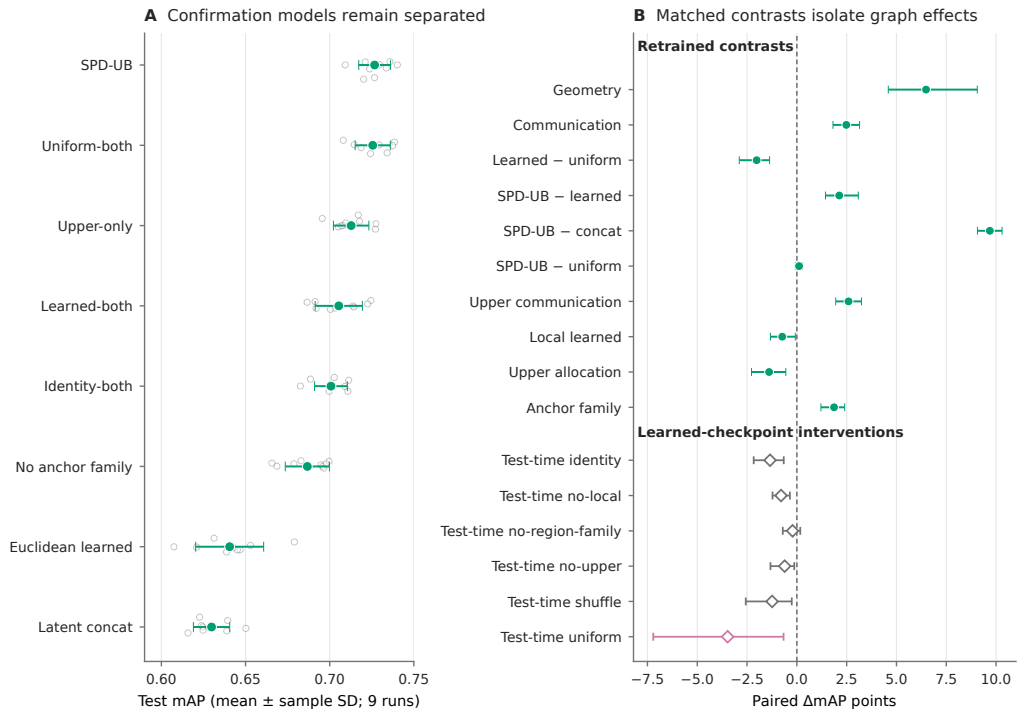


Figure 7: UTSW confirmation and structural attribution. Panel A shows nine protocol-run values together with the mean and sample SD for each confirmation model; the runs arise from three overlapping splits and three initializations, not nine external cohorts. Panel B reports paired 95% intervals for retrained structural contrasts and test-time interventions. Screening-only HeMIS, GMU, and MBT-style results ($n = 3$) are intentionally excluded from the $n = 9$ ranking.

Table B.13: UTSW subgroup and edge audits: where do support-node effects and graph stability appear?

Panel A. Anchor-family subgroup behavior

Variant/contrast	Pathology mAP	Molecular mAP
spd_cross_graph	0.7446 ± 0.0103	0.7451 ± 0.0342
spd_no_anchor_family	0.7457 ± 0.0126	0.6914 ± 0.0282
Full − no-anchor	-0.114 points	+5.372 points

Panel B. Fixed-split edge stability

Split	Scope	Spearman	Top-3 Jaccard
42	local	0.5048	0.4000
42	upper	0.5131	0.5000
43	local	0.3333	0.3000
43	upper	0.7833	0.6667
44	local	-0.0476	0.4000
44	upper	0.4476	0.6667

The anchor-family gain is concentrated in molecular retrieval and is not uniform across targets. Local stability varies by split; upper-layer stability may partly reflect fixed support blocks. Stability is not truth recovery. The three splits share test patients and are not independent external cohorts.

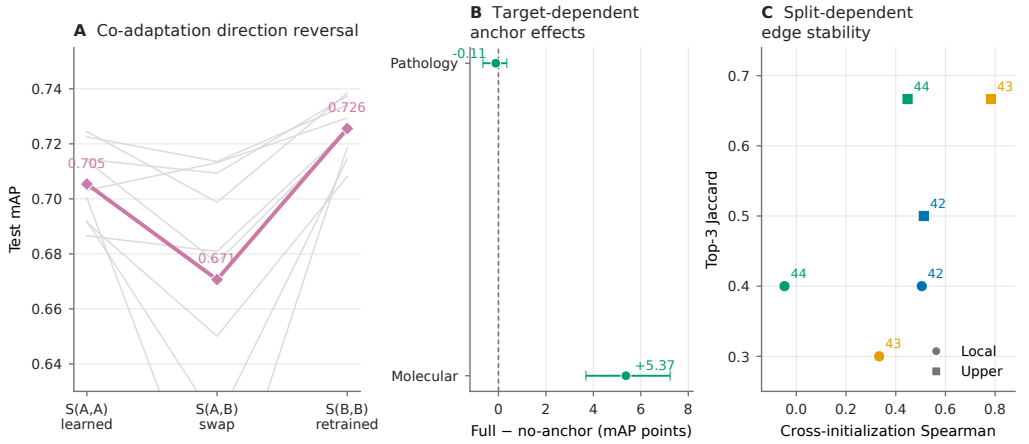


Figure 8: Co-adaptation, subgroup dependence, and edge stability on UTSW. Panel A connects the three observed train–test conditions within each matched split and initialization; $S(B, A)$ was not evaluated and is not imputed. Panel B reports full-minus-no-anchor-family paired effects, showing that the gain concentrates in molecular retrieval. Panel C labels split-specific local and upper graph stability. Stability denotes reproducibility under the fitted parameterization, not recovery of an unobserved true graph.

Table B.14: Evidence and reproducibility audit: is every reported quantity traceable to a complete protocol?

Audit item	Result	Statistical or provenance unit
Authority main completion	300/300	Training jobs
Authority attribution completion	210/210	Training jobs
Rule-only	10/10	Evaluations, not training
UTSW scheduled keys	81/81	Unique variant/split/model-seed keys
UTSW complete 50-epoch runs	81/81	History audit
Patient-level mAP recomputation	81/81	Direct score/target recomputation
Historical screening checksum match	165/165	Five critical JSON files $imes$ 33 runs
Fallback/nonfinite-query events	0	UTSW evaluation
Solver failures	0	Authority saved run records
Mixed code fingerprints	0	Protocol audit

Authority uncertainty uses five independently generated datasets; the 15 trained models are not 15 datasets. UTSW's nine runs are combinations of three overlapping splits and three initializations, not nine independent cohorts. Selection reads validation only. Summary-only retention preserves every field used here but limits new patient-level reanalysis beyond stored UTSW direct scores and targets.