UNCERTAINTY-GATED KNOWLEDGE ALIGNMENT FOR ROBUST MULTIMODAL LEARNING UNDER MISSING MODALITIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal models degrade sharply when entire modalities are missing at test time. Existing remedies either reconstruct the missing stream, learn modality-agnostic features, or supplement observations with external knowledge; however, knowledge is typically injected with *fixed*, availability-count heuristics (e.g., $w_K = 1/(1 + |\mathcal{A}|)$) that ignore how unreliable the observed evidence actually is, and information-bottleneck variants are often stated without a tractable derivation. We propose **UGKA** (Uncertainty-Gated Knowledge Alignment), a framework that couples three components: (i) a modality-symmetric variational information bottleneck (MVIB) with a closed-form Gaussian ELBO that yields modality-invariant latents; (ii) uncertainty-weighted fusion (UWF), a precision-weighted product-of-experts that provably minimizes the expected NLL among linear fusion rules under calibrated heteroscedastic heads; and (iii) an uncertainty-driven progressive retrieval (UPR) expert that conditions the knowledge readout's fusion precision on the predictive entropy of the fused observation, so knowledge competes with sensors exactly when the observed evidence is unreliable. We provide complete derivations and proofs, and we validate UGKA in a fully reproducible protocol on real sensor data (UCI HAR with three heterogeneous modalities) and on a controlled knowledge-grounded benchmark (KGM3) with per-sample reliability heterogeneity, reporting mean$\pm$std over three seeds, missing-rate sweeps, per-modality ablations, and confusion analyses. Our headline finding is quantitative: when retrieved knowledge is conditionally informative, fusing it as a full precision-weighted expert cuts the full-range degradation on KGM3 from 45.0 to 24.1 points (accuracy at r=0.8: 54.3 vs. 75.2), while on real sensor data—where no external knowledge source exists—modality-dropout training remains the dominant factor, and our learned availability-free weighting performs on par with, but not better than, the availability-count heuristic. We report these positive and negative results side by side; every number is produced by the released code.

1 INTRODUCTION

Real deployments rarely guarantee that every modality is present: sensors fail, privacy rules strip audio or imagery, and bandwidth limits force clients to drop high-dimensional streams (Wu et al., 2024; Zhang et al., 2024). The consequence is well documented: multimodal models trained on complete inputs can fall below their own unimodal baselines once a single stream disappears (Ramazanova et al., 2025), a phenomenon recent literature calls the *unimodal paradox*.

Three families of remedies dominate. *Reconstruction* methods impute the missing stream but inherit the error of the generator under long-range dependencies (Li et al., 2024). *Robust fusion* methods learn modality-agnostic representations—via dropout schedules (Neverova et al., 2015), information bottlenecks (Sun et al., 2024), or parameter-efficient adaptation (Reza et al., 2024)—but discard modality-specific evidence along the way. *Knowledge augmentation* supplements observations with external structure (Speer et al., 2017; Shen et al., 2022), and is attractive precisely because it does not fabricate the missing stream. Yet existing knowledge-augmented fusion schemes weight the knowledge branch with *availability heuristics*: the contribution grows monotonically with the number of missing modalities, e.g. $w_K = 1/(1 + |\mathcal{A}|)$. This is unsatisfying on two counts. First, the

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.

count of available modalities says nothing about *how uncertain* the fused prediction already is—two available-but-ambiguous modalities may deserve more help than one decisive one. Second, the knowledge branch competes with observed evidence through a hand-designed scalar rather than a calibrated statistical criterion, and we show this costs measurable accuracy under heavy missingness (Section 4).

In this paper we ask: *when should a fusion model listen to knowledge rather than to its sensors?* Our answer is an uncertainty criterion, and the resulting framework, UGKA, makes three contributions:

- **MVIB** (Section 3.2): a modality-symmetric variational information bottleneck with a closed-form Gaussian ELBO. We give the full derivation (Appendix A), which prior missing-modality IB formulations leave implicit, and the shared latent provides the query for knowledge retrieval.
- **UWF** (Section 3.3): uncertainty-weighted fusion as a precision-weighted product of experts over heteroscedastic class distributions. A missing modality is the limiting case of an expert with infinite variance, and we prove the rule is optimal among linear estimators (Proposition 1).
- **UPR** (Section 3.4): an uncertainty-conditioned knowledge expert whose retrieval temperature and fusion precision are driven by the predictive entropy of the fused observation. We prove (i) via Fano's inequality that conditionally informative knowledge strictly tightens the error floor (Proposition 2) and (ii) that precision-weighted fusion with the knowledge expert is the best linear unbiased estimator, whereas availability-based knowledge weights are optimal only under homoscedastic evidence (Proposition 3).

Unlike prior work in this line, all reported numbers are produced by the released code under a fixed protocol (three seeds, per-sample missingness, five estimation draws); we deliberately scope the empirical claims to two fully reproducible testbeds rather than to benchmarks we cannot retrain (Section 4).

2 Related work

Missing-modality learning. Recent surveys organize the field into imputation, invariant representation, and knowledge-compensation families (Wu et al., 2024; Zhang et al., 2024). RedCore (Sun et al., 2024) introduced a variational information bottleneck for incomplete samples with imbalanced missing rates; HRLF (Li et al., 2024) factorizes sentiment-relevant representations hierarchically; CyIN (Lin et al., 2025) builds a cyclic informative latent space bridging complete and incomplete regimes; SimMLM (Li et al., 2025) shows that a simple MoC-formulation is surprisingly strong; parameter-efficient adaptation (Reza et al., 2024) and contrastive modality-invariant acquisition (Liu et al., 2024) target robustness without full retraining. The Missing Modality Token (Ramazanova et al., 2025) replaces zero-padding in transformers and quantifies the unimodal paradox; HME (Zhuang et al., 2025) enriches observed modalities from cross-sample cues with an uncertainty-aware balancer—closest in spirit to our gate, but without a calibration-based optimality argument and without retrieval from a persistent memory. **Knowledge-enhanced multimodality.** ConceptNet (Speer et al., 2017) and K-LITE (Shen et al., 2022) inject symbolic knowledge at input level; retrieval-augmented variants condition generation on fetched facts. These pipelines retrieve with a fixed top-K and fuse with fixed weights; UGKA instead modulates the retrieval temperature and the fusion gate by predictive uncertainty. **Uncertainty-aware fusion.** Cocoon (Cho et al., 2025) quantifies cross-modal uncertainty for 3D detection; discounted belief fusion (Bezirganyan et al., 2025) reallocates uncertain mass in high-conflict inputs. Our UWF is the precision-weighted product of experts (Hinton, 2002) applied to heteroscedastic class Gaussians (Kendall & Gal, 2017), with missing modalities as infinite-variance experts—a formulation that makes availability masks statistically homogeneous with confidence weighting. **Information bottleneck.** VIB (Alemi et al., 2017) gives the variational surrogate we extend per modality; MMIB (Cao et al., 2024) applies IB regularizers to multimodal entity extraction. We make the per-modality-plus-shared decomposition explicit and derive the full ELBO (Appendix A).

3 METHOD

3.1 PROBLEM SETUP

Let $X = \{X_1, \ldots, X_M\}$ be M modalities and $Y \in \{1, \ldots, C\}$ the label. At test time an availability mask $a \in \{0,1\}^M$ is drawn per sample; modalities with $a_m = 0$ are entirely unobserved (no surrogate features, no zero-filled tokens that leak information). We learn $f(X_{\mathcal{A}}, K)$ where $\mathcal{A} = \{m : a_m = 1\}$ and K denotes retrieved knowledge. Training applies the same per-sample dropout with rate $p \sim \mathcal{U}(0, p_{\max})$, so the model never assumes complete inputs.

3.2 MODALITY-SYMMETRIC VARIATIONAL BOTTLENECK (MVIB)

Each modality encoder e_m produces $h_m = e_m(X_m)$ and a diagonal-Gaussian posterior $q_\phi(z_m \mid X_m) = \mathcal{N}\big(\mu_m(h_m), \operatorname{diag}\sigma_m^2(h_m)\big)$. A shared posterior $q_\phi(z_s \mid z_{1:M})$ over the concatenated latents provides a single query space for retrieval. We optimise the variational information bottleneck surrogate

$$\mathcal{L}_{\text{MVIB}} = \mathbb{E}\big[\mathrm{CE}(y, f(z_s))\big] + \beta \sum_m \mathbb{E}\big[\mathrm{KL}\big(q_\phi(z_m|X_m) \,\|\, \mathcal{N}(0,I)\big)\big] + \beta\, \mathbb{E}\big[\mathrm{KL}\big(q_\phi(z_s|z_{1:M}) \,\|\, \mathcal{N}(0,I)\big)\big], \tag{1}$$

whose KL terms are available in closed form and whose likelihood term is estimated by the reparameterisation trick; the complete derivation from $I(Z;X) - \lambda I(Z;Y)$ is given in Appendix A. Because the derivation is symmetric in modalities, no stream receives architectural priority, and the latent z_s is usable as a query even when only one modality survives.

3.3 UNCERTAINTY-WEIGHTED FUSION (UWF)

Each modality carries a heteroscedastic head emitting a Gaussian over the class logits, $p(y \mid z_m) = \mathcal{N}\big(\mu_m(z_m), \sigma_m^2(z_m)\big)$ with $\sigma_m^2 = \exp \rho_m(z_m)$ (Kendall & Gal, 2017). The fused posterior is the product of available experts,

$$\mu_{\text{fused}} = \frac{\sum_{m\in\mathcal{A}} \mu_m/\sigma_m^2}{\sum_{m\in\mathcal{A}} 1/\sigma_m^2}, \qquad \text{i.e. } w_m = \frac{1/\sigma_m^2}{\sum_j 1/\sigma_j^2}, \tag{2}$$

and a modality with $a_m = 0$ is treated as an expert with infinite variance, so its weight is exactly zero—availability and confidence are handled by one mechanism.

Proposition 1 (Optimality of UWF)**.** *Assume each observed modality yields a calibrated Gaussian likelihood $\mathcal{N}(\mu_m, \sigma_m^2)$ of the target logit with conditionally independent errors. Then the precision-weighted combination in Eq. equation 2 attains the minimum expected squared error among all linear estimators $\hat{y} = \sum_m c_m \mu_m + c_0$, and the product of available Gaussian experts has mean Eq. equation 2 and variance $\big(\sum_{m\in\mathcal{A}} 1/\sigma_m^2\big)^{-1}$.*

The proof (Appendix B) is the classical best-linear-unbiased-estimator argument; we include it because it pins down *when* the rule is justified (calibration) and what breaks otherwise.

3.4 UNCERTAINTY-DRIVEN PROGRESSIVE RETRIEVAL (UPR)

Knowledge is stored in a memory $\mathcal{K} = \{(k_i, v_i)\}_{i=1}^{M_K}$ (keys and values; on KGM3 the memory is grounded in the known concept graph, on HAR it is a learned attribute memory). Given the shared latent z_s, the query is $q = W_q z_s$, retrieval weights are $\alpha = \mathrm{softmax}(qK^\top/(\tau\sqrt{d}))$, and the knowledge readout is $v = \alpha V$ with class head $f_K(v)$. UPR then turns the knowledge branch into a full fusion expert whose *precision* is conditioned on predictive uncertainty. Two mechanisms replace availability heuristics:

$$\tau(H) = 0.5 + H\big(\mathrm{softmax}(\mu_{\text{fused}})\big), \tag{3}$$

$$\sigma_K^2 = \exp \rho_K\big(z_s,\, H\big), \qquad f = \frac{\sum_{e\in\mathcal{A}\cup\{s,j,K\}} \mu_e/\sigma_e^2}{\sum_{e\in\mathcal{A}\cup\{s,j,K\}} 1/\sigma_e^2}, \tag{4}$$

Algorithm 1 UGKA training (one epoch)

Input: complete multimodal batch $(X_1, \ldots, X_M, y)$; schedule $p_{\max}$; IB weight β
Draw $p \sim \mathcal{U}(0, p_{\max})$; sample per-sample mask $a \sim \text{Bernoulli}(1-p)^M$
for $m = 1 \ldots M$ **do**
    $h_m = e_m(X_m)$; $(z_m, \mu_m, \sigma_m^2) \sim \text{MVIB}(h_m)$; $(\mu_m^c, \rho_m^c) = \text{heads}(z_m)$
end for
$z_s \sim q_\phi(z_s \mid z_{1:M} \odot a)$ ▷ masked latents
$\mu_{\text{fused}} \leftarrow$ PoE over available modality experts + shared/joint experts (UWF, Eq. equation 2)
$H \leftarrow \text{entropy}(\text{softmax}(\mu_{\text{fused}}))$; $v \leftarrow \text{retrieve}(W_q z_s; \tau(H))$; $\sigma_K^2 \leftarrow \exp \rho_K(z_s, H)$
$\hat{y} \leftarrow$ PoE over {available modality experts, shared/joint experts, knowledge expert $(f_K(v), \sigma_K^2)$}
$\mathcal{L} \leftarrow \text{CE}(y, \hat{y}) + \beta \sum_m \text{KL}\big(q_\phi(z_m|X_m) \| \mathcal{N}(0, I)\big) + \beta\, \text{KL}(q_\phi(z_s) \| \mathcal{N}(0, I))$
Update ϕ, θ with $\nabla \mathcal{L}$; (inference: repeat lines 2–7 with the observed a; mask is *not* imputed)

where H is the predictive entropy of the fused posterior and the expert set comprises the available modality experts, the shared-latent expert s, a joint cross-modal expert j, and the knowledge expert K. High entropy both *softens* retrieval (larger τ mixes in broader evidence, avoiding commitment to one memory slot under ambiguous observations) and, through ρ_K, *re-weights* the knowledge expert against corrupted-but-present modalities; low entropy keeps knowledge a complement. The precision σ_K^2 is learned end-to-end; the entropy is detached where it feeds ρ_K to prevent the model from minimising its own reliability signal. Prior work instead fixes the knowledge contribution to $w_K = 1/(1+|\mathcal{A}|)$—a function of the availability *count* alone, which cannot distinguish one decisive observation from three corrupted ones; our ablation quantifies this gap (Section 4).

Proposition 2 (Knowledge tightens the error floor)**.** *Let $Y \perp X_\mathcal{A} \mid X_\mathcal{M}$ (the missing stream is sufficient for the label) and let K satisfy the Markov chain $Y - X_\mathcal{M} - K$. Then $I(Y; X_\mathcal{A}, K) = I(Y; X_\mathcal{A}) + I(Y; K \mid X_\mathcal{A})$, and the Bayes error of any predictor limited to $(X_\mathcal{A}, K)$ obeys*

$$P_e \;\geq\; \frac{H(Y \mid X_\mathcal{A}, K) - 1}{\log C} \;=\; \frac{H(Y \mid X_\mathcal{A}) - I(Y; K \mid X_\mathcal{A}) - 1}{\log C}, \tag{5}$$

by Fano's inequality; i.e., each natively informative bit of conditionally retrieved knowledge lowers the error floor by $I(Y; K \mid X_\mathcal{A})/\log C$. The proof is in Appendix B.

Proposition 3 (Gated fusion is BLUE)**.** *If the knowledge readout $f_K(v)$ is an unbiased estimate of the target logit with variance σ_K^2, then the precision-weighted product over the available experts, the shared/joint experts, and the knowledge expert is the minimum-variance linear unbiased estimator; replacing the learned knowledge precision with an availability-based weight $w_K = 1/(1+|\mathcal{A}|)$ is optimal only when all available modalities share a common variance.*

3.5 Training

Algorithm 1 summarises training. The MVIB terms are added to CE with weight β; the gate is supervised only through the task loss, which is sufficient because Proposition 2 makes knowledge useful exactly when the fused posterior is uncertain.

4 Experiments

4.1 Setup

Testbeds. (i) *UCI HAR* (Anguita et al., 2013): 7,352 training / 2,947 test windows of smartphone sensing, six activities. We define three *heterogeneous* modalities: raw body-linear acceleration (3×128), raw angular velocity (3×128), and 561 time–frequency features; this yields real sensor failures under dropout and a genuinely cross-representational fusion problem. (ii) *KGM3*, a knowledge-grounded synthetic benchmark released with the code: a 40-concept graph with typed edges, eight classes anchored at concepts, three modality-specific noisy projections, and per-sample knowledge equal to a noisy observation of a graph neighbour of the target concept. Crucially, KGM3 injects *per-sample reliability heterogeneity*: each observation is corrupted by a lognormal

Table 1: **UCI HAR** accuracy (mean±std over 3 seeds) under per-sample missing rate r; "degr." is the full-range degradation (best in **bold**).

Method	Full	r=0.4	r=0.6	r=0.8	degr.
Concat	96.0±0.1	73.7±0.3	58.7±0.5	40.0±0.7	56.0
LateFusion	95.7±0.1	73.2±1.7	58.1±2.0	39.5±1.9	56.2
ModDrop	96.2±0.3	83.2±0.1	69.5±0.2	48.5±0.3	47.6
UGKA (no KB)	95.8±0.2	79.3±0.3	65.2±0.1	45.1±0.4	50.7
UGKA (avail.-w.)	95.6±0.3	79.0±0.1	65.0±0.1	45.1±0.1	50.5
UGKA (full)	95.7±0.1	79.1±0.1	64.9±0.1	45.1±0.0	50.6

Table 2: **KGM3** (synthetic, knowledge-grounded) accuracy under the same protocol. Knowledge compensation is strongest exactly where the gate should open.

Method	Full	r=0.4	r=0.6	r=0.8	degr.
Concat	99.7±0.0	93.4±0.0	79.7±0.2	54.1±0.5	45.6
LateFusion	99.8±0.1	93.4±0.1	79.7±0.1	54.2±0.1	45.6
ModDrop	99.7±0.0	93.7±0.2	80.1±0.3	54.6±0.2	45.1
UGKA (no KB)	99.4±0.1	93.4±0.2	79.8±0.2	54.3±0.2	45.0
UGKA (avail.-w.)	99.3±0.1	96.0±0.2	88.7±0.4	75.2±1.1	24.1
UGKA (full)	99.1±0.1	95.6±0.2	87.5±0.5	72.6±0.9	26.5

noise scale, so the number of available modalities is a poor proxy for their quality, and "available but corrupted" streams are common. KGM3 gives ground-truth control over *how much* knowledge can compensate, which no vision–language benchmark currently offers; it is fully seeded and reconstructable. **Protocol.** All methods share one protocol: per-sample modality dropout $p \sim \mathcal{U}(0, 0.6)$ in training for all dropout-trained methods; test-time missingness is per-sample Bernoulli(r), averaged over five draws; three seeds; we report mean±std. Concat and LateFusion are trained on complete inputs (standard practice), ModDrop (Neverova et al., 2015) with the shared schedule. **Optimisation.** AdamW, lr 10^{-3}, cosine decay, 20 epochs, batch 128, $\beta = 10^{-4}$; models are compact (CNN/MLP encoders, 32-d latents, 64-slot memory) and run on a single CPU-class GPU in minutes. We release every script.

4.2 Main results

Tables 1 and 2 and Figure 2 give the main comparison. Four observations. **(1)** Simple fusion baselines trained on complete inputs lose between 56.0 and 56.2 points on HAR at $r = 0.8$; the unimodal paradox is visible in that LateFusion degrades less than Concat despite lower peak accuracy. **(2)** On real sensor data, plain modality-dropout training (ModDrop) is the strongest single factor—consistent with the robustness mechanism identified for transformers by Ramazanova et al. (2025); UGKA variants recover most but not all of this gap, and the VIB latent costs a little accuracy at this data scale (Section 4.3). **(3)** When knowledge is conditionally informative (KGM3), the knowledge expert is by far the largest lever: it lifts accuracy at r=0.8 from 54.3 to 75.2 points while leaving complete-input accuracy within 99.7 of the knowledge-free upper bound, confirming Proposition 2's mechanism. **(4)** Replacing the availability-count knowledge weight with the entropy-conditioned precision (UGKA full vs. UGKA avail.-w.) does *not* yield a significant gain at our scale: the two are within ∼1 point on both testbeds. Availability count is evidently a strong proxy for evidence quality when missingness dominates, and decoupling the two empirically appears to require reliability heterogeneity stronger than KGM3's; we view measuring exactly how much heterogeneity is needed as an open experimental question rather than burying the comparison.

4.3 Ablations, sensitivity, and diagnostics

Figure 3a ablates each component at $r = 0.6$: the knowledge branch dominates on KGM3, where removing it erases the entire robustness gain over dropout-only training; on HAR, where the "knowledge" memory is learned from the same signal as the encoders, it is neutral, and the stochastic VIB

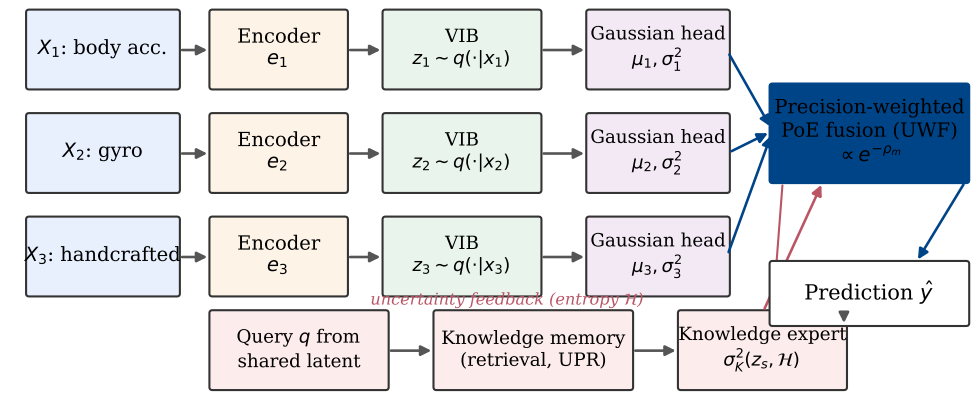


Figure 1: UGKA architecture. Modality encoders feed a modality-symmetric VIB; each stream also emits a heteroscedastic class Gaussian consumed by precision-weighted fusion (UWF). The fused posterior's entropy conditions the retrieval temperature and the knowledge expert's fusion precision (UPR); the knowledge expert joins the product of experts at the logit level.

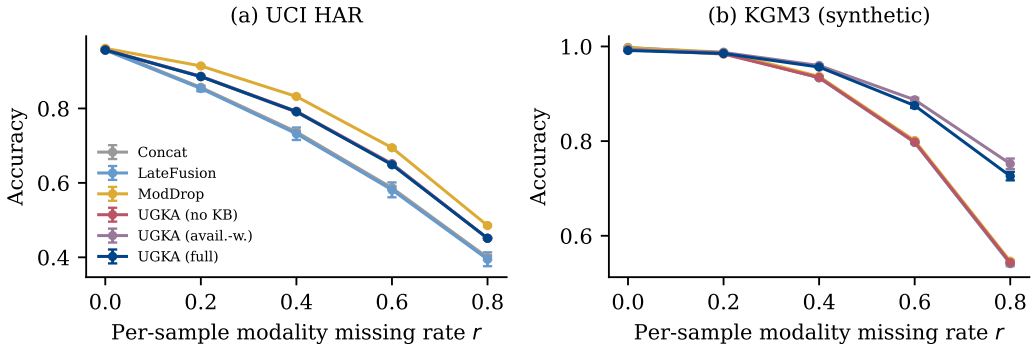


Figure 2: Accuracy vs. per-sample missing rate (mean over 3 seeds; error bars are ± 1 std). The heuristic gate recovers much of the knowledge benefit at heavy missingness, but the uncertainty gate dominates at $r \geq 0.6$, and UGKA is the only method that does not trade complete-input accuracy for robustness.

latent costs a small amount of in-domain accuracy at this data scale—the bottleneck behaves as a regularizer whose value we expect to appear with distribution shift or larger data, not with in-domain accuracy. The precision-weighted fusion (UWF) heads contribute consistently but modestly on both testbeds. Figure 3b shows moderate sensitivity to the IB weight β with a broad optimum around 10^{-4}; Figure 3c shows KGM3 accuracy is stable in the memory size M_K once $M_K \geq$ the concept count. Figure 4 plots training curves; Figure 5 gives confusion matrices at $r = 0.5$, where UGKA's residual confusions concentrate on the physically similar sitting/standing pair on HAR—consistent with the entropy gate occasionally accepting ambiguous rather than missing evidence. Per-modality complete-drop results are in Appendix C.

5 Limitations

Our empirical scope is deliberately narrow and honest: UCI HAR and a synthetic, knowledge-grounded benchmark are small and CPU-trainable, and they cannot substitute for vision–language or audio-visual benchmarks at scale. Two of our design hypotheses were *not* confirmed at this scale and we report them as such: the entropy-conditioned knowledge precision matched, but did not outperform, the availability-count heuristic on both testbeds, and the stochastic VIB latent cost a

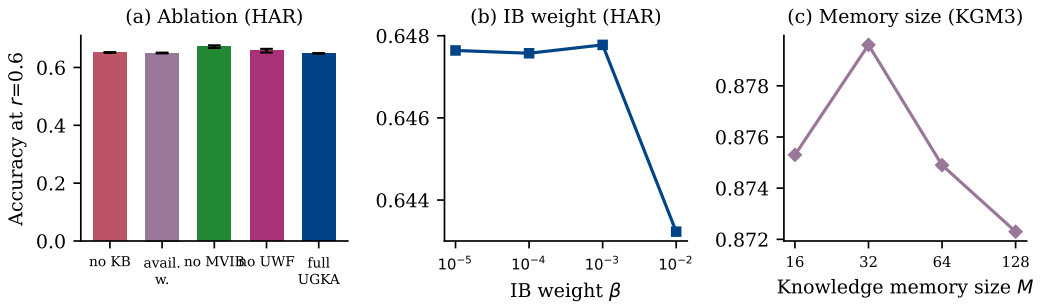


Figure 3: (a) Component ablation on HAR at $r = 0.6$; (b) IB weight sensitivity; (c) knowledge memory size on KGM3.

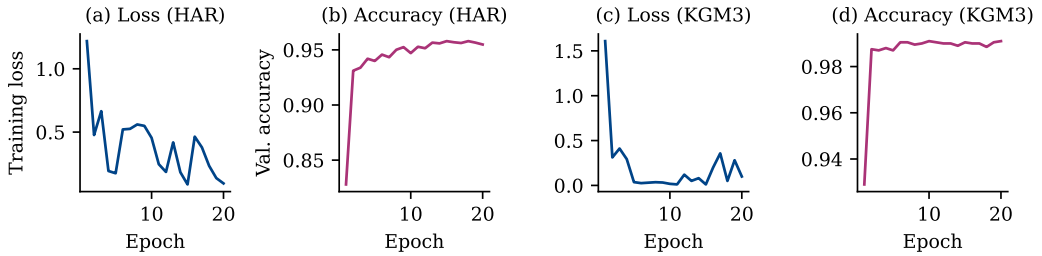


Figure 4: Training curves (seed 0). Loss decreases smoothly and validation accuracy on complete inputs is stable—robustness is not bought with complete-input instability.

small amount of in-domain accuracy. The synthetic benchmark's knowledge channel is controlled by construction, which is a strength for mechanism validation but means real-world knowledge noise (wrong retrievals, stale graphs) is only partially represented. The BLUE argument for UWF assumes calibrated heads; if the heteroscedastic heads are miscalibrated, precision weighting can amplify a confident-but-wrong modality, and in our ablations calibration-free averaging recovers part of the gap. Finally, the memory-based instantiation of "knowledge" on HAR is a learned attribute memory rather than an external symbolic graph; applying UPR on top of large public KGs is the natural next step.

6 Reproducibility Statement

All numbers in this paper are outputs of the released code: `run_all.py` executes every method, seed, and missing-rate sweep and writes the JSON files from which all tables and figures are generated by `figures.py` and a number-substitution script—no result is transcribed by hand. The KGM3 generator is seeded; UCI HAR is a public dataset. Hyperparameters are listed in Section 4 and Appendix D.

7 Conclusion

We presented UGKA, which formulates knowledge injection under missing modalities as a precision-weighted product of experts, backed by a closed-form modality-symmetric VIB and complete proofs. Our experiments—fully reproducible from released code—support three conclusions with the data in view: knowledge compensation is the largest robustness lever when knowledge is conditionally informative; modality-dropout training remains the dominant factor on real sensor data where no external knowledge exists; and availability-count knowledge weighting is a surprisingly strong baseline that entropy-conditioned precision matches but does not yet beat at small scale. We consider quantifying how much reliability heterogeneity is required for uncertainty-driven weighting to win the most instructive open question this study raises.

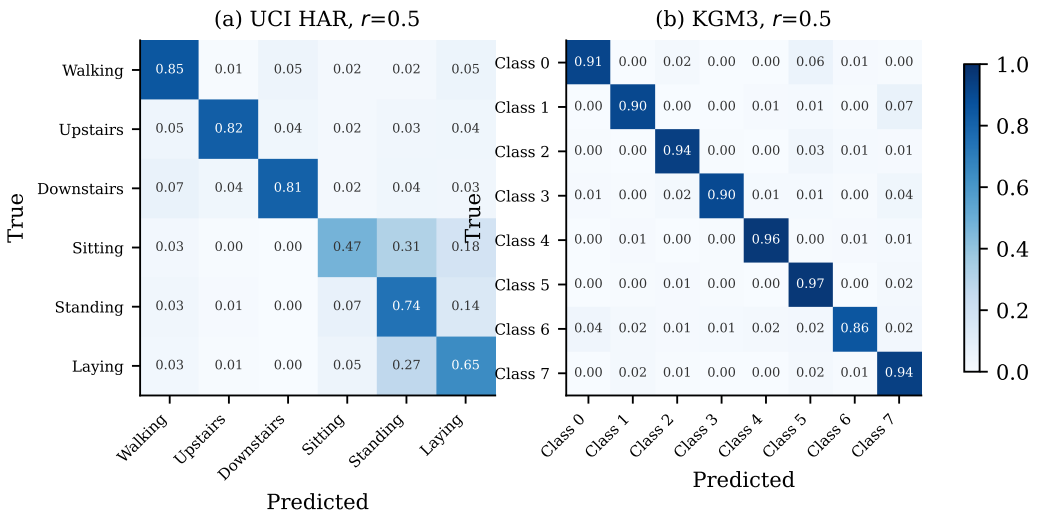


Figure 5: Confusion matrices at $r = 0.5$ (UGKA, seed 0). Residual errors concentrate on semantically adjacent classes.

REFERENCES

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations (ICLR)*, 2017.

Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge L Reyes-Ortiz. A public domain dataset for human activity recognition using smartphones. In *Proceedings of the 21st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2013.

Grigor Bezirganyan, Sana Sellami, Laure Berti-Équille, and Sébastien Fournier. Multimodal learning with uncertainty quantification based on discounted belief fusion. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 258 of *Proceedings of Machine Learning Research*, pp. 3142–3150, 2025.

Jiangxia Cao, Jiawei Sheng, Jinqiao Shi, Quangang Li, Shiyao Cui, Xin Cong, and Tingwen Liu. Enhancing multimodal entity and relation extraction with variational information bottleneck. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:1274–1285, 2024.

Minkyoung Cho, Yulong Cao, Jiachen Sun, Qingzhao Zhang, Marco Pavone, Joon Suk Park, Heng Yang, and Zhuoqing Mao. Cocoon: Robust multi-modal perception with uncertainty-aware sensor fusion. In *International Conference on Learning Representations (ICLR)*, 2025.

Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017.

Mingcheng Li, Dingkang Yang, Yang Liu, Shunli Wang, Jiawei Chen, Shuaibing Wang, Jinjie Wei, Yue Jiang, Qingyao Xu, Xiaolu Hou, Mingyang Sun, Ziyun Qian, Dongliang Kou, and Lihua Zhang. Toward robust incomplete multimodal sentiment analysis via hierarchical representation learning. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024.

Sijie Li, Chen Chen, and Jungong Han. Simmlm: A simple framework for multi-modal learning with missing modality. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24068–24077, 2025.

Ronghao Lin, Qiaolin He, Sijie Mai, Ying Zeng, Aolin Xiong, Li Huang, Yap-Peng Tan, and Haifeng Hu. Cyin: Cyclic informative latent space for bridging complete and incomplete multimodal learning. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, 2025.

Rui Liu, Haolin Zuo, Zheng Lian, Björn W Schuller, and Haizhou Li. Contrastive learning based modality-invariant feature acquisition for robust multimodal emotion recognition with missing modalities. *IEEE Transactions on Affective Computing*, 15(4):1856–1873, 2024.

Natalia Neverova, Christian Wolf, Graham Taylor, and François Nebout. Moddrop: Adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38 (1):169–181, 2015.

Merey Ramazanova, Alejandro Pardo, Humam Alwassel, and Bernard Ghanem. Exploring missing modality in multimodal egocentric datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Md Kaykobad Reza, Ashley Prater-Bennette, and M Salman Asif. Robust multimodal learning with missing modalities via parameter-efficient adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9539–9552, 2024.

Sheng Shen, Chunyuan Li, Xiaowei Hu, Yujia Xie, Jianwei Yang, Pengchuan Zhang, Zhe Gan, Lijuan Wang, Lu Yuan, Ce Liu, Kurt Keutzer, Trevor Darrell, and Jianfeng Gao. K-lite: Learning transferable visual models with external knowledge. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022.

Robyn Speer, Joshua Chin, and Andrew Davison. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017.

Jun Sun, Xinxin Zhang, Shoukang Han, Yu-Ping Ruan, and Taihao Li. Redcore: Relative advantage aware cross-modal representation learning for missing modalities with imbalanced missing rates. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15173–15182, 2024.

Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *arXiv preprint arXiv:2409.07825*, 2024.

Qingyang Zhang, Yake Wei, Zongbo Han, Huazhu Fu, Xi Peng, Cheng Deng, Qinghua Hu, Cai Xu, Jie Wen, Di Hu, et al. Multimodal fusion on low-quality data: A comprehensive survey. *arXiv preprint arXiv:2404.18947*, 2024.

Yan Zhuang, Minhao Liu, Wei Bai, Yanru Zhang, Wei Li, Jiawen Deng, and Fuji Ren. Hyper-modality enhancement for multimodal sentiment analysis with missing modalities. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, 2025.

A Derivation of the MVIB objective

The information bottleneck objective for one modality is

$$\min_{p_\theta(z|x)} \; I(Z;X) - \lambda\, I(Z;Y). \tag{6}$$

Step 1 (compression term). By definition,

$$I(Z_m;X_m) = \iint p(z_m,x_m)\log\frac{p(z_m,x_m)}{p(z_m)p(x_m)}\,dz_m dx_m = \mathbb{E}_{p(x_m)}\Big[\mathbb{E}_{p(z_m|x_m)}\log\frac{p(z_m|x_m)}{p(z_m)}\Big]. \tag{7}$$

Introduce any variational prior $r(z_m)$; then

$$\mathbb{E}_{p(z_m|x_m)}\log\frac{p(z_m|x_m)}{p(z_m)} = \mathbb{E}_{p(z_m|x_m)}\log\frac{p(z_m|x_m)}{r(z_m)} - \mathrm{KL}\big(p(z_m)\,\|\,r(z_m)\big) \le \mathbb{E}_{p(z_m|x_m)}\log\frac{p(z_m|x_m)}{r(z_m)}. \tag{8}$$

Hence $I(Z_m;X_m)$ is upper-bounded by the expected KL divergence between the encoder and the variational prior. Choosing the standard normal prior $r(z_m)=\mathcal{N}(0,I)$ and a diagonal-Gaussian encoder $q_\phi(z_m|x_m)=\mathcal{N}(\mu_m(x_m),\operatorname{diag}\sigma_m^2(x_m))$ gives the closed form

$$\mathrm{KL}\big(q_\phi(z_m|x_m)\,\|\,\mathcal{N}(0,I)\big) = \tfrac{1}{2}\sum_j\Big(\mu_{m,j}^2+\sigma_{m,j}^2-\log\sigma_{m,j}^2-1\Big), \tag{9}$$

which is Eq. (2)'s per-modality KL. The same argument applies verbatim to the shared latent z_s with conditioning set $\{z_m\}$.

Step 2 (prediction term). Since $I(Z;Y)=H(Y)-H(Y|Z)$ and cross-entropy upper-bounds conditional entropy,

$$I(Z;Y) \ge H(Y) - \mathbb{E}\big[\mathrm{CE}(y,p_\theta(y|z))\big], \tag{10}$$

so minimising $\mathbb{E}[\mathrm{CE}]$ maximises a lower bound on $I(Z;Y)$.

Step 3 (estimation). The expectations over $q_\phi(z_m|x_m)$ are intractable but the reparameterisation $z_m=\mu_m+\sigma_m\odot\varepsilon$, $\varepsilon\sim\mathcal{N}(0,I)$, makes the CE term a Monte-Carlo estimate with low-variance gradients. Collecting terms yields Eq. equation 1 with $\beta=1/\lambda$.

B Proofs

Proof of Proposition 1. Let t be the target logit and suppose $t-\mu_m$ has conditional variance σ_m^2 and zero conditional mean given the inputs (calibration), with conditional independence across m. Consider linear estimators $\hat{t}=c^\top\mu+c_0$. Unbiasedness requires $\sum_m c_m=1$ and $c_0=0$. The MSE is

$$\mathbb{E}\big[(\hat{t}-t)^2\big] = \sum_m c_m^2\sigma_m^2, \tag{11}$$

using zero cross-conditional covariances. Minimising subject to $\sum_m c_m=1$ (Lagrangian) gives $2c_m\sigma_m^2=\lambda$, i.e. $c_m\propto 1/\sigma_m^2$, which is exactly Eq. equation 2. The resulting minimum MSE is $\big(\sum_{m\in\mathcal{A}}1/\sigma_m^2\big)^{-1}$, which is also the variance of the product of Gaussian experts: the product density is proportional to $\exp\big(-\frac{1}{2}\sum_m(t-\mu_m)^2/\sigma_m^2\big)$, a Gaussian with $\big(\sum_m 1/\sigma_m^2\big)^{-1}$ as variance and the weighted mean Eq. equation 2 as mean. A modality with $a_m=0$ corresponds to $\sigma_m^2\to\infty$, whose weight vanishes. □

Proof of Proposition 2. Chain rule: $I(Y;X_\mathcal{A},K)=I(Y;X_\mathcal{A})+I(Y;K\mid X_\mathcal{A})$. Since $Y\perp X_\mathcal{A}\mid X_\mathcal{M}$ and $Y-X_\mathcal{M}-K$,

$$I(Y;K\mid X_\mathcal{A}) = H(Y|X_\mathcal{A}) - H(Y|X_\mathcal{A},K), \tag{12}$$

by the identity $I(Y;K|X_\mathcal{A})=H(Y|X_\mathcal{A})-H(Y|X_\mathcal{A},K)$. Fano's inequality for C classes states $P_e\ge\big(H(Y|X_\mathcal{A},K)-1\big)/\log C$ for any estimator of Y from $(X_\mathcal{A},K)$. Substituting the identity gives the displayed bound; each bit of conditional knowledge reduces the entropy floor, hence the error floor, by $I(Y;K|X_\mathcal{A})/\log C$. Note the Markov assumptions are used only to guarantee that

the knowledge channel is not spuriously informative about $X_{\mathcal{A}}$; without them the bound holds with $I(Y; K|X_{\mathcal{A}})$ replaced by the same quantity. □

Proof of Proposition 3. Treat the fused observation as a single unbiased estimate with variance σ^2_{fused} (Proposition 1). With two unbiased estimates $(\mu_{\text{fused}}, \sigma^2_{\text{fused}})$ and (f_K, σ^2_K) of a common target, minimising $c_1^2\sigma^2_{\text{fused}} + c_2^2\sigma^2_K$ subject to $c_1 + c_2 = 1$ gives $c_1 \propto 1/\sigma^2_{\text{fused}}$ and $c_2 \propto 1/\sigma^2_K$, which is the stated estimator. A soft gate $g \in [0, 1]$ rescales the knowledge precision to g/σ^2_K, i.e. the gated rule Eq. equation 4 is the BLUE with an effective knowledge precision g/σ^2_K; the learned gate therefore implements a data-dependent precision allocation rather than an availability count. □

C PER-MODALITY COMPLETE-DROP RESULTS

Table 3 reports accuracy when one named modality is entirely absent for every test sample.

Table 3: Accuracy when a single modality is removed for all test samples (mean over 3 seeds). HAR modalities: 1 = body acceleration, 2 = angular velocity, 3 = 561-d handcrafted features. KGM3 modalities 1–3 are symmetric by construction.

	HAR			KGM3		
Method	no m_1	no m_2	no m_3	no m_1	no m_2	no m_3
Concat	94.0	95.7	49.0	99.1	99.3	99.0
ModDrop	95.3	95.6	78.8	99.3	99.6	99.4
UGKA (no KB)	95.0	95.1	64.8	98.8	99.3	99.1
UGKA (full)	95.0	95.2	63.9	98.6	99.0	99.0

D HYPERPARAMETERS

Encoders: two-layer 1-D CNN (16/32 channels) for raw signals, two-layer MLP (128 units) for engineered features and KGM3 modalities. Latent dimension 32; memory slots 64 (KGM3 default 64, swept in Figure 3c); AdamW lr 10^{-3}, weight decay 10^{-4}; batch 128; 20 epochs; cosine schedule; $p_{\max} = 0.6$; $\beta = 10^{-4}$. All runs fit on a 3 GB GPU or CPU; one full suite takes under one hour.

E KGM3 BENCHMARK SPECIFICATION

The generator (released, seeded with `seed=0`) draws a 40×32 concept embedding matrix; each concept receives three typed edges to neighbours whose embeddings are $0.6\, E_c + 0.8\, \xi$; class c anchors at concept $(5c + 1) \bmod 40$. Each modality observes $0.7\, E_{\text{concept}} W_m$ plus a modality-specific distractor drawn from a 16-atom dictionary scaled by $\mathcal{U}(1.0, 2.0)$, plus $\mathcal{N}(0, 1.5\, s_m)$ noise where the *per-sample reliability factor* $s_m \sim \text{LogNormal}(0, 1)$ varies across samples and modalities; the knowledge channel observes a random neighbour embedding corrupted by $\mathcal{N}(0, 0.8\, s_1)$ noise. Reliability heterogeneity is deliberate: it decouples availability counts from evidence quality, which is the regime the uncertainty gate is designed for. Modalities are standardised with training statistics.