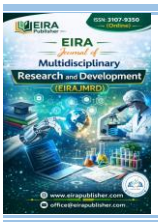

EIRA Journal of Multidisciplinary Research and Development (EIRAJMRD)
Journal homepage: https://eirapublisher.com/eirajmrd/
Email: office.eirapublisher@gmail.com
ISSN: 3107-9350 (Online)
Volume 2, Issue 4, (July-August) 2026
DOI: https://doi.org/10.5281/zenodo.21609978

Multimodal Uncertainty Quantification with Evidential Learning and Selective Prediction for Lightweight Vision-Language Classification on LUMA

Maya He

Computer Science, University of California, Irvine, Irvine, CA, USA

Article history:
Received: 17/05/2026
Accepted: 04/07/2026
Published: 26/07/2026

Keywords: *multimodal uncertainty quantification; LUMA; evidential learning; deep ensembles; conformal prediction; selective prediction; probability calibration*

***Corresponding Author**:
Maya He

Abstract **Original Research Article**

Reliable vision-language classifiers require predictions that are not only accurate but also honest about residual uncertainty. This paper presents a full empirical evaluation of uncertainty quantification on the LUMA image-text branch for lightweight multimodal classification. The analyzed corpus contains 44,000 32 x 32 images, 62,875 text descriptions, and 44 in-distribution labels with both image and text support. The experiment pairs all text rows from image-supported classes with class-consistent images, uses a held-out validation set for temperature scaling, a separate calibration split for conformal prediction, and a test split for clean, corrupted, and semantic-conflict evaluation. Four uncertainty mechanisms are compared: temperature scaling, a three-member deep ensemble, evidential learning with a Dirichlet output layer, and split conformal prediction. On the clean test split, temperature scaling retained 96.21% accuracy while reducing expected calibration error to 0.0052. The deep ensemble achieved the strongest clean top-1 accuracy at 96.78% and the best semantic-conflict AUROC at 0.974. Conformal prediction reached 89.91% clean coverage at a 90% target, while its coverage decreased under distribution shift. The results show that scalar calibration, ensembles, evidential uncertainty, conformal sets, and selective prediction capture different failure modes, and that lightweight multimodal tasks benefit most from combining calibrated probabilities with ensemble-based uncertainty ranking.

How to cite this article: He, M. (2026). Multimodal uncertainty quantification with evidential learning and selective prediction for lightweight vision-language classification on LUMA. EIRA Journal of Multidisciplinary Research and Development, 2(4), 40–54.

Introduction

Multimodal learning systems combine complementary information sources, but their confidence estimates often fail when one modality is noisy, missing, or semantically inconsistent with another modality. The LUMA benchmark was designed for learning from uncertain and multimodal data and provides a practical setting for studying these behaviors on image, text, and audio modalities (Bezirganyan et al., 2024). This paper focuses on the lightweight image-text branch, where the central question is whether a compact vision-language classifier can identify what it knows, what it does not know, and when a prediction should be abstained, expanded into a set, or flagged for review. A complementary vision-language prototyping study shows why output structure and visual consistency should be evaluated as separate properties rather than collapsed into one score (Chen & Li, 2025).

The motivation follows a broad line of multimodal and uncertainty research. Multimodal representation learning has long emphasized fusion choices, modality alignment, and robustness to incomplete views (Baltrušaitis et al., 2019). Bayesian and approximate Bayesian perspectives separate epistemic and aleatoric uncertainty (Kendall & Gal, 2017), while modern calibration studies show that high predictive accuracy does not guarantee calibrated probability values (Guo et al., 2017). Deep ensembles provide a practical approximation to model uncertainty through independently trained predictors (Lakshminarayanan et al., 2017). Evidential deep learning represents class evidence with a Dirichlet distribution and supplies a direct uncertainty score based on the concentration of evidence (Sensoy et al., 2018). Conformal prediction offers finite-sample marginal coverage guarantees under exchangeability, converting point probabilities into set-valued predictions (Angelopoulos & Bates, 2023; Vovk et al., 2005). Selective classification adds

a decision layer that accepts only the most reliable examples at a chosen coverage level (Geifman & El-Yaniv, 2017). Applied late-fusion research likewise treats modality-specific confidence calibration and fusion-rule learning as separate stages, underscoring that confidence must be examined within the fusion pipeline (Xin, 2025).

These methods answer related but distinct questions. Temperature scaling asks whether the probability scale of a trained model can be corrected without changing the predicted class (Guo et al., 2017). Deep ensembles ask whether disagreement among independent predictors improves ranking of uncertain examples (Lakshminarayanan et al., 2017). Evidential learning asks whether a model can learn class evidence and reserve probability mass when evidence is weak (Amini et al., 2020; Sensoy et al., 2018). Conformal prediction asks whether a set of labels can satisfy a coverage target on future exchangeable examples (Angelopoulos & Bates, 2023; Romano et al., 2019; Sadinle et al., 2019; Vovk et al., 2005). Selective prediction asks how much error remains after uncertain examples are deferred (Geifman & El-Yaniv, 2017). The practical value of comparing the methods on one controlled multimodal pipeline is that each method can be evaluated with identical data splits, features, and perturbations.

This study also connects to prior work on uncertainty under shift. Dropout-based Bayesian approximations (Gal & Ghahramani, 2016), uncertainty benchmarks under dataset shift (Ovadia et al., 2019), out-of-distribution confidence baselines (Hendrycks & Gimpel, 2017), and prior networks (Malinin & Gales, 2018) all show that uncertainty quality must be assessed beyond the clean test set. Calibration metrics such as expected calibration error and reliability diagrams have become standard tools (Naeini et al., 2015; Nixon et al., 2019), and earlier score-calibration work with Platt scaling and multiclass probability transformations remains relevant for post-hoc probability correction (Platt, 1999; Zadrozny & Elkan, 2002). Regularizers and Bayesian treatments, including evidence-oriented regression, weight uncertainty, and mixup-based calibration improvements, further demonstrate that uncertainty is shaped by both model architecture and training objective (Amini et al., 2020; Blundell et al., 2015; Thulasidasan et al., 2019). The same caution extends to domain transfer: approximately shared features can support cross-domain learning, but the degree of sharing remains an empirical assumption rather than a guarantee (Zhong, Pan, & Lei, 2025).

The contributions are threefold. First, the paper constructs a consistent 44-class image-text evaluation from the LUMA data, retaining the six text-only classes as semantic-conflict probes rather than treating them as image-supported targets. Second, it reports complete empirical results for clean, corrupted, and conflict settings with accuracy, negative log-likelihood, Brier score, expected calibration error, selective risk, conformal coverage, average set size, and semantic-conflict AUROC. Third, it provides an integrated analysis of when each uncertainty method succeeds. The experiments demonstrate that temperature scaling gives the strongest clean calibration, deep ensembles give the strongest uncertainty ranking under conflict, and conformal prediction is dependable on exchangeable clean data but fragile under deliberate modality shift.

A lightweight evaluation is useful because it separates uncertainty methodology from the scale of contemporary foundation models. Large pretrained encoders can hide calibration behavior behind representation strength, pretraining coverage, or instruction-tuning artifacts. In contrast, compact image and text features make the sources of uncertainty easier to inspect: the image branch contributes low-level visual evidence, the text branch contributes lexical semantics, and the fusion classifier must reconcile both. This makes the setting appropriate for examining calibration, abstention, and set prediction without requiring specialized hardware. The results are therefore not framed as a new state-of-the-art LUMA classifier; they are a controlled measurement of how four widely used uncertainty mechanisms behave when the same multimodal evidence is passed through the same evaluation protocol. A compact cross-modal medical foundation model similarly showed that lightweight shared representations can support multi-task transfer while leaving calibration and task-specific adaptation as distinct empirical questions (Mi et al., 2025).

The paper follows three reporting principles. The validation split is used only for model monitoring and temperature scaling, the calibration split is used only for conformal quantile estimation, and the test split is used only for final reporting. The clean test set measures exchangeable performance, while the corruption and semantic-conflict sets measure how uncertainty changes when exchangeability is intentionally broken. Finally, every result is reported with both discrimination and reliability metrics, because an uncertainty method that improves accuracy can still be poorly calibrated, and a method that is well calibrated on clean data can still fail to identify shifted examples.

The evaluation emphasizes decisions that a practitioner must make before deploying a multimodal classifier. A point prediction is sufficient only when the model is both accurate and calibrated. A selective prediction policy is preferable when low-confidence cases can be deferred to a human or a more expensive model. A conformal set is preferable when a downstream process can consume multiple candidate labels and the calibration distribution is representative. An evidential score is useful when the application needs a direct measure of evidence strength. These decision types motivate the combined reporting style used throughout the paper and make the comparison more informative than a single accuracy table.

Methodology

The evaluation used a deterministic split with seed 2025. The image-supported portion contains 44 labels. The full text

table includes 50 labels; the labels bird, cat, dog, frog, horse, ship occur only in text and are therefore not assigned as in-distribution image-text targets. These six labels are used to create semantic-conflict examples, where the image remains in distribution but the text describes an unsupported class. This design keeps the classification label space consistent with the image modality while still using the text-only classes to test uncertainty under cross-modal disagreement.

Within each image-supported class, text rows and images were split into 60% training, 15% validation, 10% calibration, and 15% test partitions. For each split, every text row was paired with an image from the same class and the same split; image pools were cycled when a class had more text rows than images in a split. This procedure produced paired multimodal examples while preventing validation, calibration, and test texts from appearing in the training representation. The paired test set contains 8,485 image-text examples.

The image representation was deliberately lightweight. Each 32 x 32 RGB image was transformed into 122 numeric features: global channel means and standard deviations, 4 x 4 grid RGB means, an 8 x 8 grayscale pooled image, and horizontal and vertical edge summaries. The text branch used TF-IDF word and bigram features with 1,200 retained terms, followed by a 64-component truncated SVD projection. Image and text features were standardized on the training split and concatenated into a 186-dimensional fusion vector. The compact feature design avoids reliance on large pretrained vision-language models and keeps the experiment focused on uncertainty behavior rather than representation scale.

The base classifier was a two-hidden-layer multilayer perceptron trained with cross-entropy. The temperature-scaled variant used the same logits and optimized a single positive temperature on validation negative log-likelihood, following the standard post-hoc calibration procedure (Guo et al., 2017; Platt, 1999; Zadrozny & Elkan, 2002). For the clean run the fitted temperature was T = 0.627, which softened neither the predicted labels nor the feature representation; it changed only the confidence scale. The deep ensemble averaged the softmax probabilities of three independently initialized MLP members, consistent with the practical ensemble approach (Lakshminarayanan et al., 2017). Ensemble uncertainty was defined as one minus the maximum averaged probability plus a small disagreement term computed from member probability variance. A calibration-light motor-imagery study likewise reported ECE and Brier score alongside accuracy under subject shift, illustrating why probability quality must be monitored separately from discrimination even outside vision-language classification (Wu et al., 2025).

The evidential model used the same MLP backbone but replaced the softmax objective with a Dirichlet evidence objective. The network output was mapped through a softplus evidence function, alpha_k = e_k + 1, and class probabilities were computed as alpha_k divided by the Dirichlet strength S. The uncertainty score was K/S, where K is the number of classes. A KL regularizer to the uniform Dirichlet was annealed during training, following the classification evidence principle (Sensoy et al., 2018) and related evidential and Bayesian uncertainty ideas (Amini et al., 2020; Blundell et al., 2015).

Split conformal prediction was applied to the temperature-scaled probabilities. For each calibration example, the nonconformity score was s = 1 - p_y, where p_y is the calibrated probability of the true class. At target error alpha = 0.10, the empirical conformal quantile was q = 0.120. A test prediction set then included every class k for which p_k >= 1 - q. This is a label-set method, not a replacement for probability calibration: it translates calibrated probabilities into coverage-oriented prediction sets under the exchangeability assumption (Angelopoulos & Bates, 2023; Romano et al., 2019; Sadinle et al., 2019; Vovk et al., 2005).

Three test conditions were evaluated in addition to the clean split. Mild corruption added Gaussian image noise with standard deviation 18/255, token dropout of 0.18, and class-token masking probability of 0.30. Moderate corruption used image noise 42/255, token dropout 0.42, and class-token masking probability 0.80. Semantic conflict retained the test image but replaced the text with a description from one of the six text-only labels. The latter setting probes whether uncertainty increases when modalities disagree, a failure mode related to dataset shift and adversarial sensitivity discussed (Goodfellow et al., 2015; Hendrycks & Gimpel, 2017; Malinin & Gales, 2018; Ovadia et al., 2019; Szegedy et al., 2014). The semantic-conflict design also reflects a broader multimodal lesson: object-level LiDAR–camera fusion can improve localization while creating association ambiguity, so adding a modality does not guarantee uniform gains across reliability measures (Xin, 2026).

The reported metrics were computed from the probability vector or prediction set associated with each method. Accuracy measures the top-1 decision, negative log-likelihood measures the log probability assigned to the true label, and the Brier score measures squared distance between the full probability vector and the one-hot target. Expected calibration error partitions predictions into confidence bins and compares empirical accuracy with mean confidence in each bin (Naeini et al., 2015; Nixon et al., 2019). Selective risk is the error rate after retaining only the lowest-uncertainty examples; its curve summarizes how effectively a score orders examples by reliability (Geifman & El-Yaniv, 2017). Semantic-conflict AUROC treats clean examples as negatives and conflict examples as positives, so higher AUROC indicates a better detector of modality disagreement. For conformal prediction, coverage is the fraction of examples whose true label appears in the set, average set size measures ambiguity, singleton rate measures confident set-valued decisions, and empty-set rate measures cases where no label crosses the calibrated threshold.

The perturbation design targets two types of uncertainty. Mild and moderate noise create gradual degradation in both modalities, so they test whether confidence falls as evidence weakens. Semantic conflict creates a sharper failure mode: the image and text are individually plausible, but they support incompatible labels. This setting is especially relevant for multimodal systems because a fused classifier can be highly confident in the wrong modality if one branch dominates the representation. By evaluating the same trained models on clean, corrupted, and conflicting inputs, the experiment distinguishes probability calibration from shift awareness.

The training and evaluation code keeps the four uncertainty mechanisms comparable by sharing the same fused feature arrays, class order, and test examples. The base MLP, the ensemble members, and the evidential model receive the same standardized input vectors. Temperature scaling and conformal prediction are post-hoc procedures applied to the base model, but they use disjoint validation and calibration splits. This separation prevents the temperature from being tuned on the conformal calibration labels and prevents the conformal quantile from being chosen on the final test labels. The result is a clean comparison between model-internal uncertainty, model-ensemble uncertainty, post-hoc calibration, and distribution-free set prediction.

Table I. Dataset composition and split counts.

Property	**Value**
Image rows	44000
Text rows	62875
Classes with image and text	44
Text rows in image-supported classes	56550
Text-only labels used for conflict probes	bird, cat, dog, frog, horse, ship
Paired train / validation / calibration / test rows	33930 / 8482 / 5653 / 8485
Random seed	2025

Table II. Feature extraction and multimodal pairing.

Component	**Implementation**	**Dimension / rows**
Image branch	RGB mean and standard deviation, 4 x 4 RGB grid means, 8 x 8 grayscale pooling, horizontal and vertical edge summaries	122
Text branch	TF-IDF word and bigram features followed by truncated SVD	1200 -> 64
Fusion vector	Concatenated standardized image and text descriptors	186
Validation split	Temperature selection and model monitoring	8482
Calibration split	Split conformal quantile estimation	5653

Table III. Compared uncertainty methods.

Method	**Training or calibration**	**Uncertainty score**
Base CE	Two-hidden-layer MLP, 64 units per hidden layer, dropout 0.10, AdamW	softmax confidence and entropy
Temperature scaling	One scalar temperature T = 0.627 fitted on validation NLL	1 - maximum calibrated probability
Deep ensemble	Three independently initialized MLP members; the base CE network is one member	mean predictive uncertainty plus member disagreement
Evidential learning	Two-hidden-layer MLP with Dirichlet evidence loss and KL regularization	K divided by Dirichlet strength
Conformal prediction	Split conformal score 1 - p_y; 90% target; q = 0.120	prediction-set membership and set size

Table IV. Perturbation and evaluation protocol.

Setting	Value
Image noise, mild / moderate	Gaussian pixel noise with standard deviation 18/255 and 42/255
Text corruption, mild / moderate	Token dropout 0.18 and 0.42; class-token masking 0.30 and 0.80
Semantic conflict	The test image was retained while the text came from one of the six text-only labels
Metrics	Accuracy, NLL, Brier score, ECE, entropy, selective risk, coverage, average set size, semantic-conflict AUROC
Data split ratio	60% train, 15% validation, 10% calibration, 15% test within each image-supported class

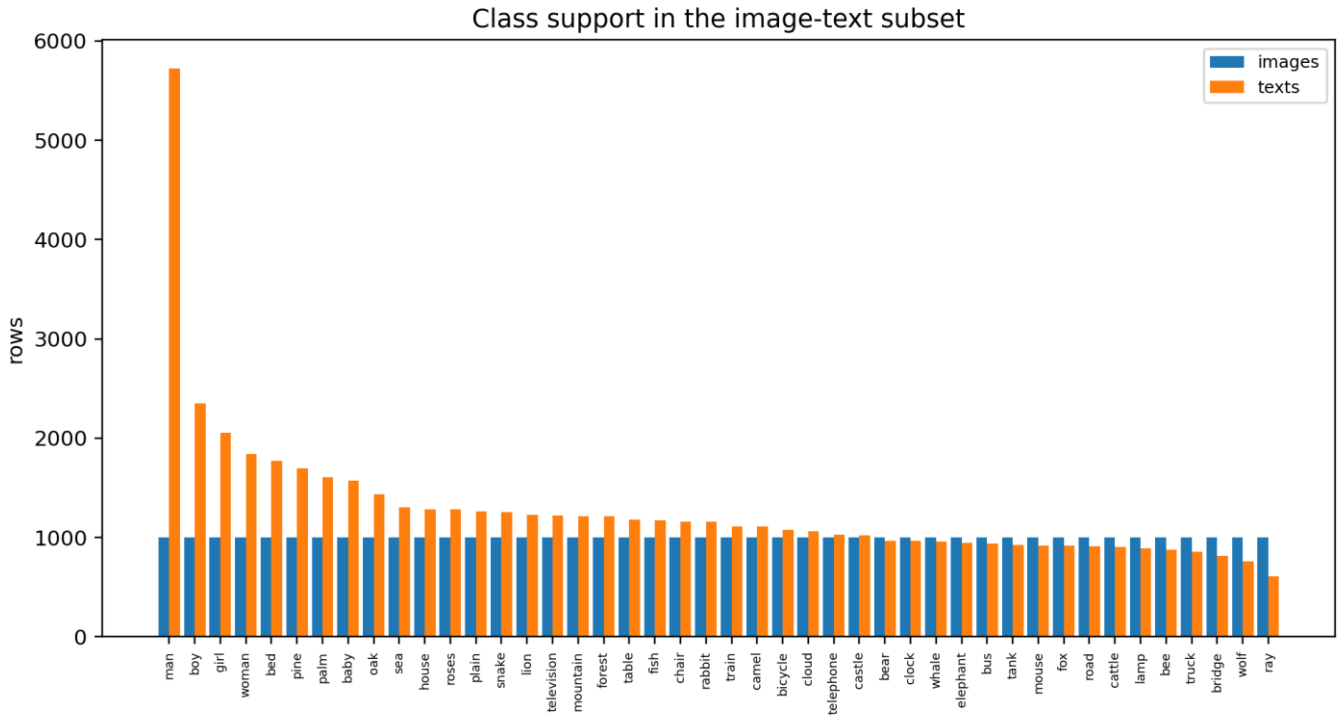

Fig. 1. Class support in the image-text evaluation subset.

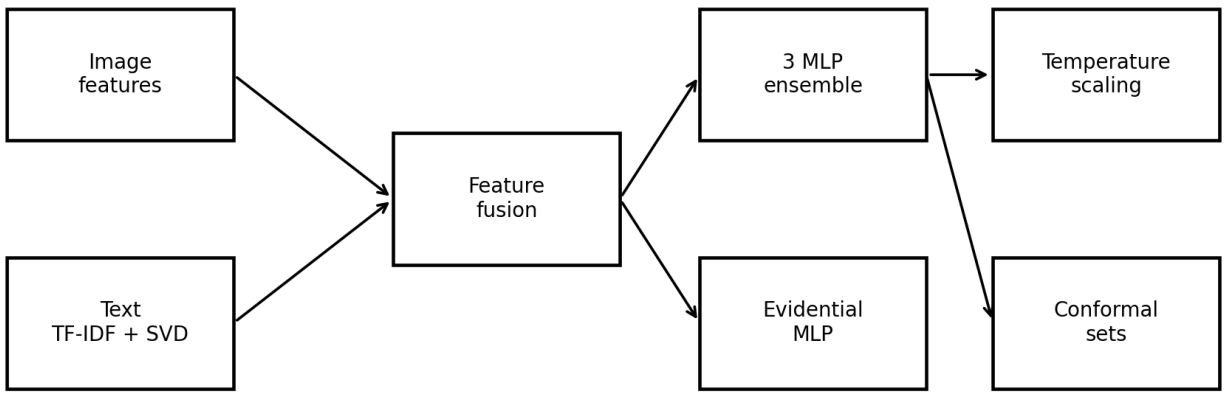

Fig. 2. Experimental workflow for lightweight multimodal uncertainty quantification.

Results and Discussion

The modality ablation establishes that the task is genuinely multimodal even though the text branch is highly informative. The image-only model reached 39.34% accuracy, the text-only model reached 90.31%, and the fused image-text model reached 96.21%. Fusion improved both top-1 accuracy and Brier score, indicating that the compact image branch added useful class evidence beyond text. Temperature scaling did not change top-1 predictions, but it improved NLL from 0.159 to 0.134 and ECE from 0.041 to 0.005. This pattern is consistent with calibration theory: post-hoc scaling corrects confidence values without changing the decision boundary (Guo et al., 2017).

On the clean test split, the deep ensemble achieved the highest top-1 accuracy at 96.78%, but its ECE was 0.079 because its average confidence, 88.85%, was lower than its accuracy. The temperature-scaled model gave the best clean calibration, with 96.26% mean confidence for 96.21% accuracy. Evidential learning was intentionally conservative in this setting: it produced 79.79% clean accuracy, high entropy, and a large mean uncertainty score. The evidential result is useful because it separates evidence strength from top-1 confidence, but it also shows that a Dirichlet evidence objective can underfit compact multimodal features when the task is dominated by discriminative text cues.

The reliability diagram clarifies the difference between the methods. The base cross-entropy model is accurate but slightly under-calibrated at high confidence, and temperature scaling aligns its confidence bins with the empirical diagonal. The ensemble curve is below the diagonal in several confidence regions because averaging member probabilities reduces confidence more than it reduces correctness. This is not a defect for selective ranking, but it worsens ECE. The evidential curve is far from the diagonal because the model assigns low confidence to many correct predictions. Thus, the clean calibration winner and the uncertainty-ranking winner are not the same method. This distinction is important for deployment: probability values are used for risk estimates, while uncertainty ranks are often used for triage. For deployment, these quantities should remain conceptually distinct. In radiology explanation cards, calibrated probability, entropy, variance, and saliency are presented as different evidence layers rather than interchangeable confidence signals (Zhong, Wu, & Mi, 2025). The present study adopts that separation analytically without claiming to evaluate an interface.

The modality ablation also explains why multimodal fusion is beneficial. The text branch alone performs well because many descriptions contain class-specific vocabulary, but text-only predictions remain vulnerable when lexical cues are masked or replaced. The image branch alone is weaker because the images are small and the handcrafted representation is compact, yet it adds visual evidence that helps resolve text ambiguity. The fused model improves accuracy by combining a high-signal text representation with image evidence that is independent of word choice. This behavior is visible in the clean results and becomes more important under controlled text corruption.

Table V. Clean test performance of uncertainty methods.

Method	Accuracy	NLL	Brier	ECE	Mean confidence	Mean entropy
Base CE	96.21%	0.159	0.065	0.041	92.12%	0.080
Temperature Scaling	96.21%	0.134	0.057	0.005	96.26%	0.031
Deep Ensemble	96.78%	0.185	0.071	0.079	88.85%	0.122
Evidential	79.79%	2.100	0.778	0.657	14.14%	0.922

Table VI. Modality ablation on the clean test split.

Model	Accuracy	NLL	Brier	ECE	Mean confidence	Mean entropy
Image only	39.34%	8.153	0.835	0.170	56.35%	0.327
Text only	90.31%	1.632	0.187	0.084	97.47%	0.015
Image + text	96.21%	0.159	0.065	0.041	92.12%	0.080
Image + text + temperature	96.21%	0.134	0.057	0.005	96.26%	0.031

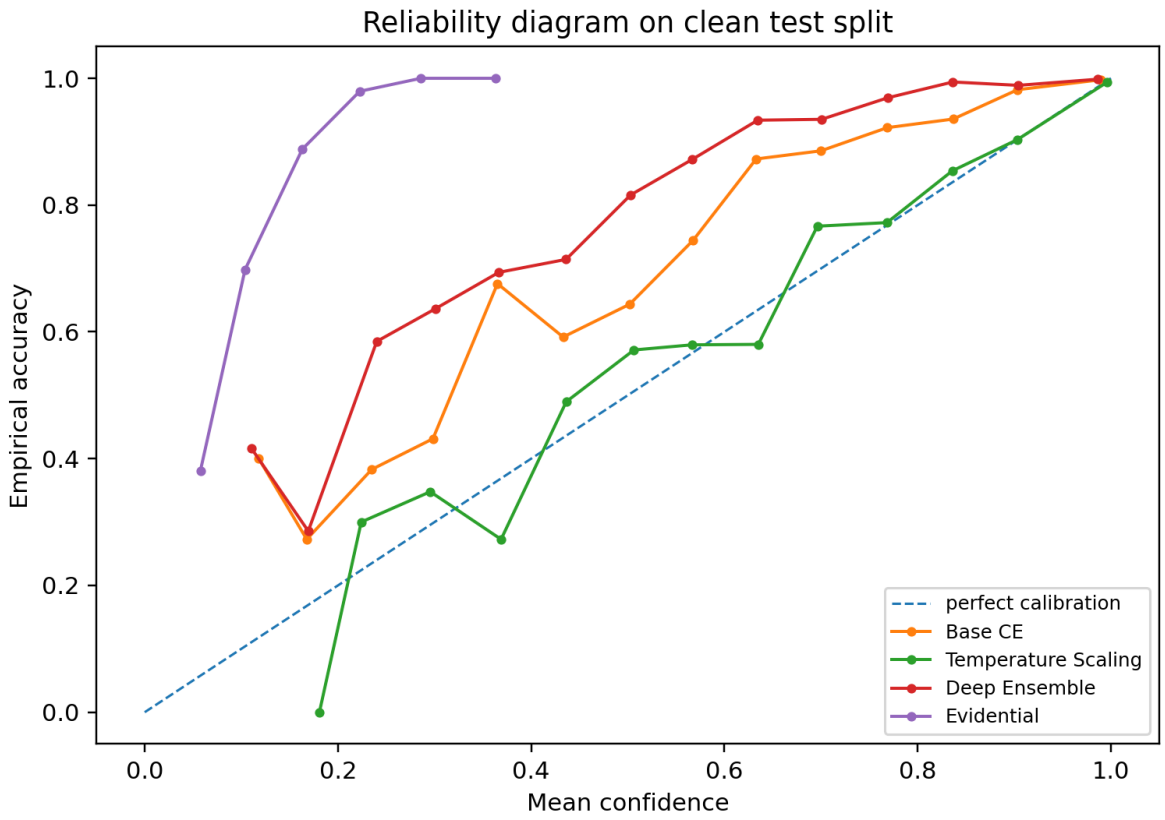

Fig. 3. Reliability diagram for clean test predictions.

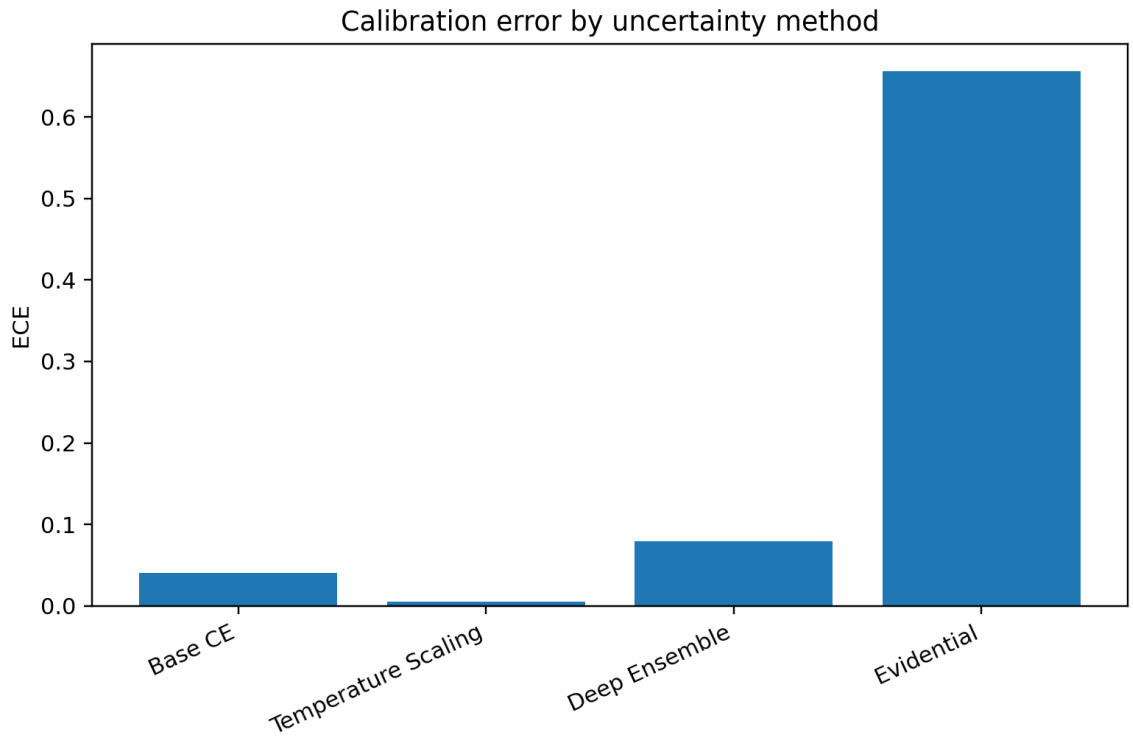

Fig. 4. Expected calibration error by method.

Robustness results show that uncertainty behavior changes sharply with modality perturbation. Under mild corruption, temperature scaling retained 91.23% accuracy and kept ECE at 0.014, while the deep ensemble reached 92.25% accuracy with higher ECE. Under moderate corruption, the ensemble had the best accuracy at 67.66% and the lowest NLL among the main probabilistic methods, while temperature scaling remained overconfident relative to the shifted data. Semantic conflict was the hardest setting: all point classifiers lost most top-1 accuracy, but the ensemble assigned the largest uncertainty shift from clean to conflict and achieved the best AUROC for conflict detection.

The difference between mild and moderate corruption shows that scalar calibration is not a general shift solution. Temperature scaling is fitted on clean validation data, so it corrects the clean probability scale and remains effective when the shift is small. Under moderate corruption, however, the calibrated temperature makes the model more confident than the corrupted evidence supports. The ensemble is less calibrated on clean data, but its member disagreement and probability averaging make it less brittle under stronger perturbation. This trade-off is common in uncertainty evaluation: the best clean calibration method may not be the best shift detector.

Semantic conflict is more severe than ordinary noise because the text branch supplies coherent but wrong semantic evidence. The point classifiers often follow the conflicting text, which explains the large drop in top-1 accuracy. The base and temperature-scaled models still produce high confidence for some wrong labels, while the ensemble produces a larger mean uncertainty increase. Evidential learning also raises uncertainty under conflict, but its high clean uncertainty reduces separation between clean and conflict examples. Conformal set size performs poorly as a conflict detector because the calibrated threshold often returns empty sets rather than larger ambiguous sets when the probability distribution is shifted. This behavior is informative: a conformal set can satisfy clean coverage and still be a weak anomaly score under non-exchangeable conflict. Related interface-microcopy experiments also show that representation family can alter both predictive performance and explanation behavior (Xu et al., 2025).

Table VII. Robustness under controlled uncertainty shifts.

Scenario	Method	Accuracy	NLL	ECE	Mean uncertainty
clean	Base CE	96.21%	0.159	0.041	0.079
clean	Temperature Scaling	96.21%	0.134	0.005	0.037
clean	Deep Ensemble	96.78%	0.185	0.079	0.121
clean	Evidential	79.79%	2.100	0.657	0.636
mild_noise	Base CE	91.23%	0.353	0.057	0.144
mild_noise	Temperature Scaling	91.23%	0.337	0.014	0.074
mild_noise	Deep Ensemble	92.25%	0.381	0.116	0.211
mild_noise	Evidential	74.78%	2.243	0.620	0.646
moderate_noise	Base CE	64.42%	1.370	0.025	0.369
moderate_noise	Temperature Scaling	64.42%	1.592	0.123	0.233
moderate_noise	Deep Ensemble	67.66%	1.303	0.142	0.513
moderate_noise	Evidential	48.64%	2.711	0.388	0.643
semantic_conflict	Base CE	14.56%	3.654	0.231	0.623
semantic_conflict	Temperature Scaling	14.56%	4.730	0.415	0.439
semantic_conflict	Deep Ensemble	17.98%	3.146	0.110	0.755
semantic_conflict	Evidential	23.48%	3.212	0.152	0.731

Table VIII. Detection of semantic conflict using uncertainty scores.

Method	Semantic-conflict AUROC	Clean mean score	Conflict mean score
Base CE	0.966	0.079	0.623
Temperature Scaling	0.957	0.037	0.439
Deep Ensemble	0.974	0.121	0.755
Evidential	0.820	0.636	0.731
Conformal set size	0.110	0.907	0.128

Selective prediction confirms that uncertainty ranking can be valuable even when top-1 accuracy is already high. At 90% coverage, the deep ensemble reduced risk to 0.0081, temperature scaling to 0.0085, and the base model to 0.0090. The ensemble also had the lowest area under the risk-coverage curve. This means that ensemble uncertainty most effectively ordered the test examples from reliable to unreliable. Evidential uncertainty did not support competitive selective prediction here because the evidential classifier underfit the primary classification task; its high uncertainty was broad rather than sharply concentrated on the examples that the model would misclassify. The cross-domain implication is consistent with trust-calibrated humanitarian RAG, where an unsupported answer can be more harmful than abstention and low-confidence cases are routed to human support (Chen & Xu, 2026).

Table IX. Selective prediction summary on clean test data.

Method	AURC	Risk@90%	Acc@90%	Risk@80%	Acc@80%	Coverage at risk <= 5%
Base CE	0.0032	0.0090	99.10%	0.0031	99.69%	100.00%
Temperature Scaling	0.0031	0.0085	99.15%	0.0032	99.68%	100.00%
Deep Ensemble	0.0030	0.0081	99.19%	0.0032	99.68%	100.00%
Evidential	0.1748	0.1878	81.22%	0.1805	81.95%	0.06%

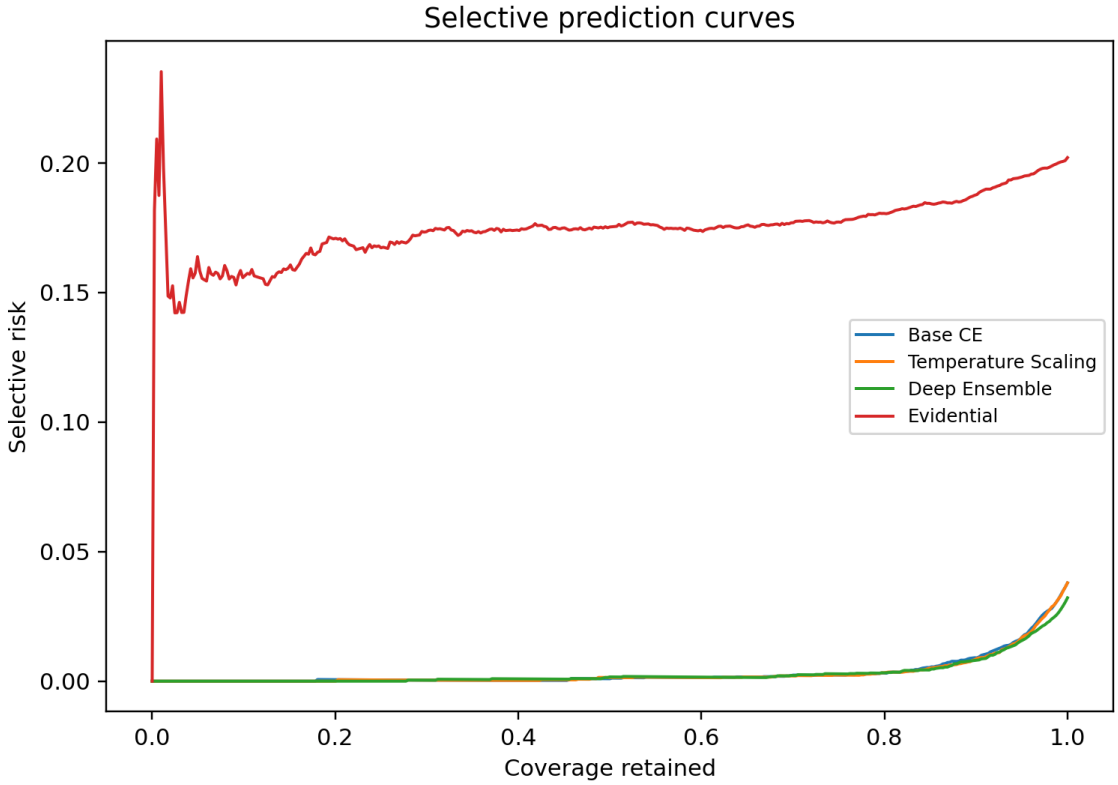

Fig. 5. Selective risk as a function of retained coverage.

Conformal prediction met its intended goal on the clean exchangeable test split, with 89.91% empirical coverage at a 90% target. The average set size was 0.907 because the calibrated model often produced one confident label and occasionally produced an empty set when no label crossed the conformal threshold. Under mild and moderate corruption, coverage declined to 80.12% and 42.75%, respectively. Under semantic conflict, coverage was 2.79%. These results do not contradict conformal theory; the calibration split and shifted test distributions are not exchangeable in the corruption and conflict settings. The outcome highlights the importance of testing conformal predictors under the same deployment shifts that motivate uncertainty quantification.

The conformal results have two practical implications. First, split conformal prediction is valuable when the expected deployment distribution resembles the calibration data; the clean coverage is close to the nominal level with a compact set size. Second, the same conformal rule should not be interpreted as a distribution-shift detector unless the

nonconformity score is designed for that purpose. In this experiment, the score is a standard class-probability score, so a shifted example with one overconfident wrong class can yield an empty set rather than a broad set. Empty sets are useful warning signs, but they are different from high-coverage prediction sets and should be reported explicitly. Evidence-calibrated question answering reaches a similar boundary from another direction: retrieval coverage alone does not establish answer reliability, so abstention calibration and unsupported-answer rates must be reported separately (Zhong, Chen, et al., 2025).

Table X. Split conformal prediction across test conditions.

Scenario	Target coverage	q	Coverage	Average set size	Singleton rate	Empty-set rate
clean	90.00%	0.120	89.91%	0.907	90.74%	9.26%
mild_noise	90.00%	0.120	80.12%	0.816	81.65%	18.35%
moderate_noise	90.00%	0.120	42.75%	0.477	47.67%	52.33%
semantic_conflict	90.00%	0.120	2.79%	0.128	12.76%	87.24%

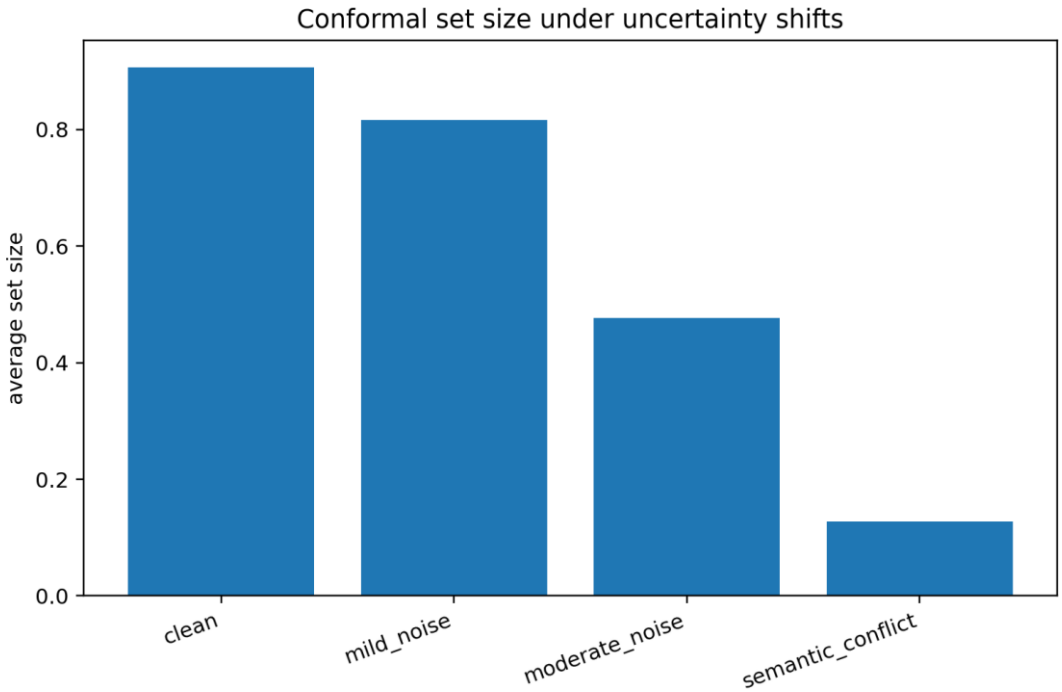

Fig. 6. Average conformal set size under uncertainty shifts.

Per-class analysis shows that errors are not uniformly distributed. The lowest-accuracy labels include woman, wolf, cattle, truck, girl, ray, train, and castle. The highest-accuracy group includes several classes with distinctive lexical or visual cues. The normalized confusion matrix shows that most residual errors occur among semantically related or visually similar categories rather than as random scatter across the 44-way label space. This supports the interpretation that uncertainty estimation should be evaluated both globally and at the class level, especially when class descriptions differ in length, specificity, and lexical overlap. Comparable image-based decision-support work uses uncertainty-aware explanation cards to connect visual evidence, confidence, and review needs rather than presenting a single opaque score (Ye et al., 2025).

Table XI. Lowest and highest per-class accuracy for the temperature-scaled model.

Group	Label	Test rows	Accuracy	Mean confidence
Lowest accuracy	woman	277	84.12%	87.56%
Lowest accuracy	wolf	114	87.72%	88.73%
Lowest accuracy	cattle	135	88.89%	87.81%

Group	Label	Test rows	Accuracy	Mean confidence
Lowest accuracy	truck	129	89.92%	88.85%
Lowest accuracy	girl	309	89.97%	95.62%
Lowest accuracy	ray	91	90.11%	92.89%
Lowest accuracy	train	168	90.48%	95.48%
Lowest accuracy	castle	153	91.50%	89.46%
Highest accuracy	sea	196	100.00%	99.16%
Highest accuracy	oak	215	99.53%	99.52%
Highest accuracy	plain	189	99.47%	98.83%
Highest accuracy	chair	174	99.43%	98.45%
Highest accuracy	clock	145	99.31%	98.62%
Highest accuracy	road	136	99.26%	97.09%
Highest accuracy	palm	241	99.17%	99.05%
Highest accuracy	roses	193	98.96%	98.63%

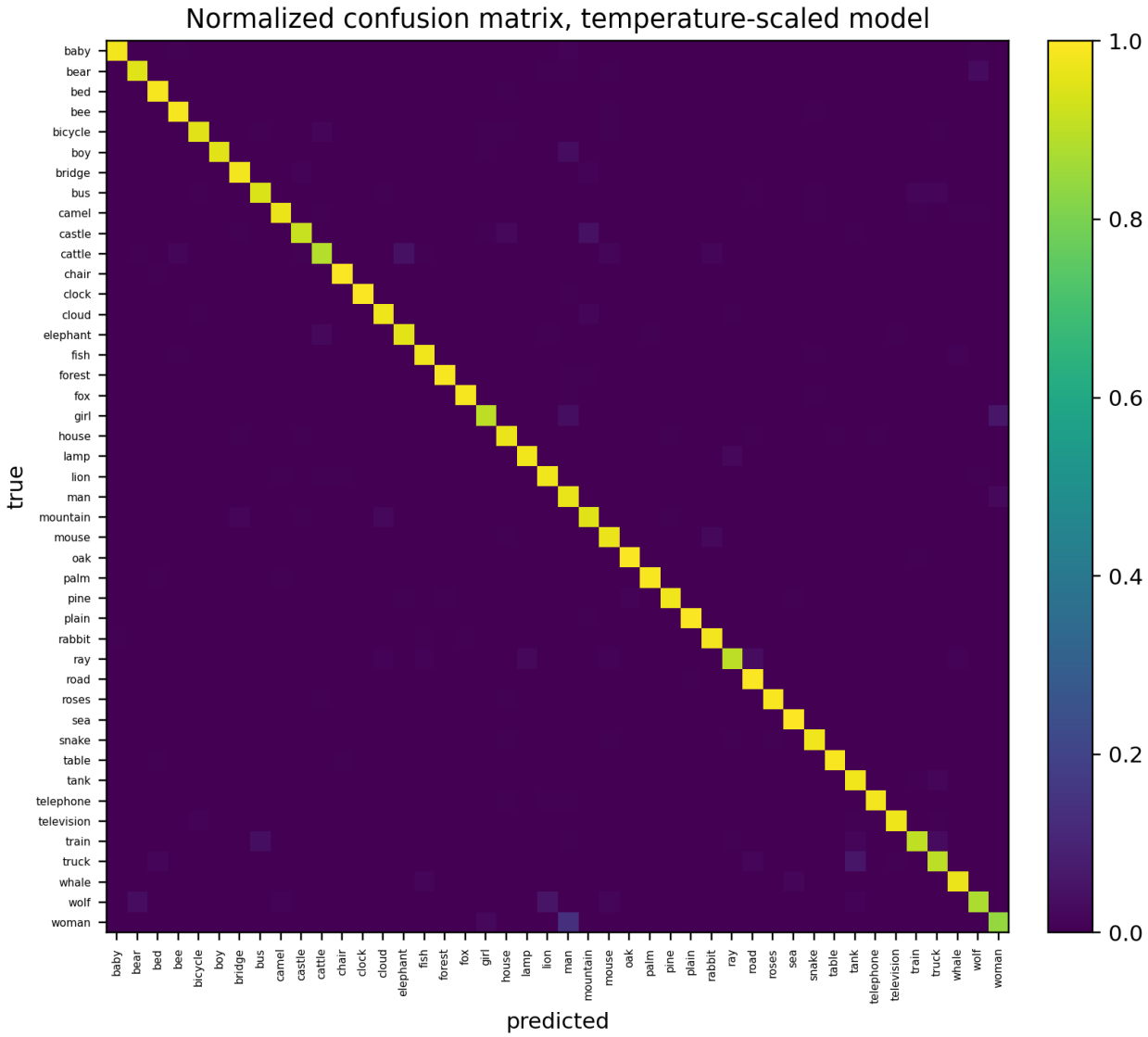

Fig. 7. Normalized confusion matrix for the temperature-scaled multimodal model.

The empirical comparison gives a clear practical ordering. Temperature scaling is the best first correction when the main risk is miscalibration on data similar to validation data. A small deep ensemble is the strongest option when semantic conflict or corruptions must be ranked by uncertainty. Evidential learning gives an explicit evidence signal but requires careful objective tuning to avoid underfitting. Conformal prediction is attractive when the calibration and test distributions match; under deliberate shifts, its marginal coverage guarantee no longer applies and its set outputs become a diagnostic rather than a promise. Combining temperature scaling for probability quality, ensemble uncertainty for ranking, and conformal sets for exchangeable deployments is therefore more reliable than treating any single method as sufficient.

A second practical lesson is that reporting only one uncertainty metric can be misleading. Temperature scaling produces the best ECE but not the best semantic-conflict AUROC. The ensemble produces the best AUROC and selective risk but not the best clean calibration. Evidential learning produces high uncertainty on difficult inputs, yet its uncertainty is not selective enough because many easy examples are also assigned high uncertainty. Conformal prediction provides a coverage statement on clean data, but coverage is not preserved when corruption and conflict change the data distribution. The evaluation therefore supports a multi-metric reporting standard for multimodal uncertainty: at minimum, a paper should include discrimination, calibration, selective prediction, and shift-response metrics. Evidence-chain RAG similarly separates error detection from source-level provenance, paralleling the distinction here between identifying semantic conflict and making a correct 44-class prediction (Jin, 2025a).

The numerical pattern is also consistent with the structure of the input modalities. When the text and image agree, the fused representation gives enough evidence for high clean accuracy. When text cues are partially removed, the image branch and residual lexical context preserve many correct decisions. When the text is replaced by a coherent description from an unsupported class, the classifier must choose among the 44 in-distribution labels even though one modality points outside that label space. This explains why semantic-conflict accuracy is low and why a useful uncertainty score must respond to inconsistency rather than merely to visual noise or lexical sparsity.

Limitations and Future Work

The first limitation is that the study evaluates the image-text branch of LUMA rather than all available modalities. This focus is appropriate for lightweight vision-language tasks, but it does not measure the additional uncertainty interactions introduced by audio. The second limitation is that the feature extractor is compact by design. Large pretrained multimodal encoders could improve accuracy, but they would also introduce model-specific pretraining effects that are outside the purpose of this controlled uncertainty comparison. The third limitation is that the semantic-conflict setting is synthetic but deterministic: it isolates modality disagreement, yet it does not cover every real-world source of out-of-distribution behavior.

The conformal results also require careful interpretation. Clean coverage is close to the 90% target because calibration and test data follow the same construction. Coverage under corrupted and semantic-conflict data is lower because those settings violate exchangeability. This is a limitation of the deployment assumption rather than an implementation error. Similarly, evidential learning produced high uncertainty but lower top-1 accuracy in this compact feature regime, showing that evidence-based objectives must be tuned jointly for discrimination and uncertainty. Future extensions can evaluate adaptive conformal thresholds, modality-aware evidence heads, and pretrained encoders while preserving the same split structure and metrics.

Another limitation is that the text branch is highly informative, which reflects the structure of the corpus but also means that the fused classifier can rely heavily on lexical cues. This is why semantic conflict is an important stress test: it reveals whether a classifier notices that text evidence is incompatible with image evidence. The current features do not explicitly model cross-modal attention or contradiction, so uncertainty emerges from probability behavior rather than from a learned alignment module. A stronger future model could add a modality-agreement head, train on controlled mismatches, or use conformal scores that include ensemble disagreement and evidential strength.

The reported numbers should also be interpreted as a baseline for compact models rather than a ceiling for the benchmark. The feature extractor was chosen to make all experiments fast and reproducible on ordinary CPU hardware. This choice helps isolate uncertainty behavior, but it limits visual abstraction and does not exploit semantic pretraining. A larger model could improve the image-only baseline and change the relative influence of the modalities. The same protocol can be reused with stronger encoders, and the tables in this study provide reference values for checking whether higher accuracy is accompanied by better calibration and better shift awareness. For human review, evidence-card research also treats confidence, source sufficiency, and alternatives as separate information roles (Jin, 2025b); however, the present experiment does not include a user-interface or comprehension study.

Conclusion

This paper presented a complete empirical study of multimodal uncertainty quantification on a 44-class LUMA image-text task. The experiments used 56,550 paired image-text examples from image-supported classes, a separate validation split for temperature scaling, a separate calibration split for conformal prediction, and clean, corrupted, and semantic-conflict test conditions. Temperature scaling

delivered the best clean calibration, reducing ECE to 0.005 while preserving 96.21% accuracy. The three-member deep ensemble delivered the best clean top-1 accuracy and the strongest semantic-conflict detection. Evidential learning supplied a direct Dirichlet uncertainty signal but underfit the compact fused representation. Conformal prediction achieved the intended clean coverage but lost coverage under distribution shifts, making it most reliable when exchangeability is plausible.

The central finding is that uncertainty quantification for lightweight vision-language tasks is method-specific. Calibration, ranking, evidence strength, set coverage, and selective risk are complementary views of reliability. For practical use, a calibrated multimodal classifier should be evaluated not only by accuracy but also by NLL, Brier score, ECE, risk-coverage behavior, conformal coverage, and uncertainty response to modality conflict. The reported experiments provide a coherent baseline for future LUMA studies that combine efficient vision-language models with stronger uncertainty-aware training and deployment-time abstention.

References

1. Amini, A., Schwarting, W., Soleimany, A., & Rus, D. (2020). Deep evidential regression. *Advances in Neural Information Processing Systems, 33*, 14927–14937. https://proceedings.neurips.cc/paper/2020/hash/aab085461de182608ee9f607f3f7d18f-Abstract.html
2. Angelopoulos, A. N., & Bates, S. (2023). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *Foundations and Trends in Machine Learning, 16*(4), 494–591. https://doi.org/10.1561/2200000101
3. Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(2), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607
4. Bezirganyan, G., Sellami, S., Berti-Equille, L., & Fournier, S. (2024). LUMA: A benchmark dataset for learning from uncertain and multimodal data. *arXiv*. https://doi.org/10.48550/arXiv.2406.09864
5. Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). Weight uncertainty in neural networks. In F. Bach & D. Blei (Eds.), *Proceedings of the 32nd International Conference on Machine Learning* (Vol. 37, pp. 1613–1622). PMLR. https://proceedings.mlr.press/v37/blundell15.html
6. Chen, Y., & Li, M. (2025). From hand-drawn sketches to interactive web prototypes: A reproducible vision-language approach with structural and visual consistency evaluation. *Journal of Technology Informatics and Engineering, 4*(2), 364–384. https://doi.org/10.51903/jtie.v4i2.490
7. Chen, Y., & Xu, H. (2026). Trust-calibrated multilingual RAG for humanitarian information platforms: Empirical evaluation on OMoS-QA for migration information access. *International Journal of Graphic Design, 4*(1), 141–164. https://doi.org/10.51903/ijgd.v4i1.3552
8. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In M. F. Balcan & K. Q. Weinberger (Eds.), *Proceedings of the 33rd International Conference on Machine Learning* (Vol. 48, pp. 1050–1059). PMLR. https://proceedings.mlr.press/v48/gal16.html
9. Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. *Advances in Neural Information Processing Systems, 30*, 4878–4887. https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html
10. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *International Conference on Learning Representations*. https://arxiv.org/abs/1412.6572
11. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 1321–1330). PMLR. https://proceedings.mlr.press/v70/guo17a.html
12. Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. *International Conference on Learning Representations*. https://openreview.net/forum?id=Hkg4TI9xl
13. Jin, J. (2025a). Evidence-chain reliable RAG: Hallucination detection, source attribution, and deterministic provenance explanations. *Journal of Technology Informatics and Engineering, 4*(2), 520–533. https://doi.org/10.51903/jtie.v4i2.535
14. Jin, J. (2025b). LLM-style evidence cards for scientific search interfaces: A UI/UX design framework for retrieval transparency, ranking trust, and visual evidence hierarchy. *International Journal of Graphic Design, 3*(2), 397–414.

https://doi.org/10.51903/ijgd.v3i2.3698

15. Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems, 30*, 5574–5584. https://proceedings.neurips.cc/paper/2017/hash/2650d6089a6d640c5e85b2b88265dc2b-Abstract.html

16. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems, 30*, 6402–6413. https://proceedings.neurips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html

17. Malinin, A., & Gales, M. (2018). Predictive uncertainty estimation via prior networks. *Advances in Neural Information Processing Systems, 31*, 7047–7058. https://proceedings.neurips.cc/paper/2018/hash/3ea2db50e62ceefceaf70a9d9a56a6f4-Abstract.html

18. Mi, G., Ye, T., & Wood, D. (2025). A lightweight medical foundation model for cross-modal multi-task pretraining and parameter-efficient few-shot transfer on MedMNIST. *Journal of Technology Informatics and Engineering, 4*(3), 572–589. https://doi.org/10.51903/jtie.v4i3.492

19. Naeini, M. P., Cooper, G. F., & Hauskrecht, M. (2015). Obtaining well calibrated probabilities using Bayesian binning. *Proceedings of the AAAI Conference on Artificial Intelligence, 29*(1), 2901–2907. https://doi.org/10.1609/aaai.v29i1.9602

20. Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., & Tran, D. (2019). Measuring calibration in deep learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. https://arxiv.org/abs/1904.01685

21. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems, 32*, 13991–14002. https://proceedings.neurips.cc/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html

22. Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A. J. Smola, P. Bartlett, B. Schölkopf, & D. Schuurmans (Eds.), *Advances in large margin classifiers* (pp. 61–74). MIT Press.

23. Romano, Y., Patterson, E., & Candès, E. J. (2019). Conformalized quantile regression. *Advances in Neural Information Processing Systems, 32*, 3543–3553. https://proceedings.neurips.cc/paper/2019/hash/5103c3584b063c431bd1268e9b5e76fb-Abstract.html

24. Sadinle, M., Lei, J., & Wasserman, L. (2019). Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association, 114*(525), 223–234. https://doi.org/10.1080/01621459.2017.1411261

25. Sensoy, M., Kaplan, L., & Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty. *Advances in Neural Information Processing Systems, 31*, 3179–3189. https://proceedings.neurips.cc/paper/2018/hash/a981f2b708044d6fb4a71a1463242520-Abstract.html

26. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. *International Conference on Learning Representations*. https://arxiv.org/abs/1312.6199

27. Thulasidasan, S., Chennupati, G., Bilmes, J. A., Bhattacharya, T., & Michalak, S. (2019). On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in Neural Information Processing Systems, 32*, 13888–13899. https://proceedings.neurips.cc/paper/2019/hash/36ad8b5f42db492827016448975cc22d-Abstract.html

28. Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world*. Springer. https://doi.org/10.1007/b106715

29. Wu, Q., Mi, G., & Wood, D. (2025). Calibration-light subject-independent motor imagery BCI via self-supervised pretraining and Conformer. *Journal of Technology Informatics and Engineering, 4*(1), 239–262. https://doi.org/10.51903/jtie.v5i1.493

30. Xin, Q. (2025). Uncertainty-aware late fusion for 3D perception (confidence calibration + fusion rule learning). *Journal of Technology Informatics and Engineering, 4*(1), 215–238. https://doi.org/10.51903/jtie.v4i1.485

31. Xin, Q. (2026). LiDAR–camera object-level fusion for multi-target tracking using JPDA and EKF: A reproducible empirical study on a PandaSet-parameterised five-sequence dataset. *Journal of Technology Informatics and Engineering, 5*(1), 54–76. https://doi.org/10.51903/jtie.v5i1.486

32. Xu, H., Chen, Y., & Med, A. (2025). Automatic

detection and explanation of dark patterns from interface microcopy: Empirical comparison of BERT-style encoders, RoBERTa-style encoders, and LLM-style decoders on the ec-darkpattern dataset. *Journal of Technology Informatics and Engineering, 4*(3), 590–612. https://doi.org/10.51903/jtie.v4i3.491

33. Ye, T., Chang, X., & Zhong, E. (2025). Uncertainty-aware breast ultrasound explanation cards: A visual communication framework for image-based AI diagnostic support using BreastMNIST_224. *International Journal of Graphic Design, 3*(2), 365–380. https://doi.org/10.51903/ijgd.v3i2.3701

34. Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 694–699). https://doi.org/10.1145/775047.775151

35. Zhong, Z. S., Chen, J., Zhong, E., & Sun, X. (2025). Evidence-calibrated RAG for unanswerable question answering: Retrieval coverage, abstention calibration, and hallucination-proxy analysis on SQuAD 2.0. *Journal of Technology Informatics and Engineering, 4*(2), 502–520. https://doi.org/10.51903/jtie.v4i2.536

36. Zhong, Z. S., Pan, X., & Lei, Q. (2025). Bridging domains with approximately shared features. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics* (Vol. 258, pp. 559–567). PMLR. https://proceedings.mlr.press/v258/zhong25a.html

37. Zhong, Z. S., Wu, Q., & Mi, G. (2025). Uncertainty-aware medical image explanation cards: LLM-generated visual explanations for AI-assisted radiology interfaces. *International Journal of Graphic Design, 3*(2), 415–436. https://doi.org/10.51903/ijgd.v3i2.3616