Received: 07 June 2026, Revised: 03 July 2026, Accepted: 06 July 2026, Online: 03 September 2026

DOI: https://dx.doi.org/10.55708/js0509001

Retrieval-Augmented Reliability-Aware Selective Inference for Visual Classification

Pratheswaran Hariharan[1], **Haiping Xu**[1], **Donghui Yan**[*,2]

[1]Department of Computer and Information Science, University of Massachusetts, Dartmouth, MA 02747, USA

[2]Department of Mathematics and Program in Data Science, University of Massachusetts, Dartmouth, MA 02747, USA

Email(s): phariharan@umassd.edu (P. Hariharan), hxu@umassd.edu (H. Xu), dyan@umassd.edu (D. Yan)

[*]Corresponding author: Donghui Yan, dyan@umassd.edu

ABSTRACT: Multimodal large language models (MLLMs) can generate fluent visual responses even when the underlying visual prediction is weak, ambiguous, or incorrect. This work presents a retrieval-augmented reliability-aware selective inference method that evaluates the strength and consistency of visual evidence before a prediction is communicated through a downstream multimodal response. A pretrained ResNet-50 encoder extracts normalized visual embeddings, and FAISS retrieves the top-k reference images from an ImageNet-100 evidence database. Prediction reliability is assessed using retrieval similarity, class-support agreement, evidence margin, entropy-based uncertainty, and an aggregate reliability score. A decision gate then determines whether the prediction should be accepted, presented cautiously, or rejected through abstention or fallback. The selected decision is used to control the final user-facing response generated by the model. Experiments on ImageNet-100 show that the proposed method improves the accuracy of returned predictions from 85.84% to 88.88% at 89.04% coverage. The accepted error rate decreases from 14.16% to 11.12%, corresponding to a 3.04-percentage-point absolute reduction and a 21.48% relative reduction. Expected Calibration Error decreases from 7.40% to 5.87%, while high-reliability wrong predictions decrease from 263 to 226. These results demonstrate the potential of retrieval-derived evidence and selective decision gating as a post-hoc reliability mechanism for controlling visual predictions before they are incorporated into multimodal responses. The current evaluation is conducted in a controlled visual-classification setting and does not constitute a complete assessment of free-form MLLM hallucination.

KEYWORDS: Multimodal large language models, visual classification, selective prediction, retrieval-augmented inference, reliability estimation, uncertainty estimation, abstention.

1. Introduction

Multimodal Large Language Models (MLLMs) have significantly advanced vision-language understanding by combining visual encoders with large language models to interpret images and generate natural-language responses. These systems support task such as image description, visual question answering, cross-model retrieval, and multimodal reasoning. Their ability to communicate visual interpretations in fluent language makes them increasingly useful in user-facing applications. However, fluent and confident responses do not necessarily indicate that the underlying visual interpretation is correct. When an image is ambiguous, degraded, visually similar to several categories, or insufficiently represented in the available reference data, an incorrect prediction may still be communicated in a convincing manner.

A major source of this reliability problem is the always-answering behavior of many visual and multimodal systems. Such systems are generally designed to return a prediction for every input, even when the available evidence is weak or conflicting. Internal confidence scores and softmax probabilities are commonly used as indicators of certainty, but they may be poorly calibrated and may not correspond to an incorrect prediction, particularly for visually similar classes, uncertain samples, degraded images, or inputs that differ from the reference distribution.

Selective inference provides a practical alternative to this behavior by separating prediction generation from prediction acceptance. Instead of treating every prediction as equally trustworthy, a selective system evaluates whether sufficient evidence is available before returning an answer. Predictions supported by string and consistent evidence can be accepted, partially supported by string and consistent evidence can be accepted, partially supported predictions can be communicated cautiously, and weak or conflicting cases can be rejected through abstention or fallback. This introduces a reliability-coverage trade off in which the system may sacrifice a limited amount of coverage to improve the correctness and trustworthiness of the outputs it returns.

Retrieval-augmented inference offers an interpretable source of external visual evidence for this decision process. Rather than evaluating a query image in isolation, a system can retrieve visually similar reference examples and examine whether the retrieved neighborhood consistently

supports a particular class. However, retrieval alone does not guarantee correctness. Retrieved examples may be only weakly similar, divided across competing classes, or systematically unrelated to the true content of the query. Therefore, both the strength and the internal consistency of the retrieved evidence must be evaluated before it is used to justify a final prediction.

This work presents a retrieval-augmented reliability-aware selective inference method for visual classification. A pretrained ResNet-50 encoder is used to extract normalized visual embeddings, and FAISS is used to retrieve the top-k nearest reference examples from an ImageNet-100 evidence database. The retrieved neighborhood is evaluated using top similarity, mean similarity, class-support agreement, evidence margin, entropy-based uncertainty, and an aggregate reliability score. These signals are then processed by a decision gate that determines whether the prediction should be accepted, presented cautiously, or rejected through abstention or fallback. The method therefore does not automatically return the class with the highest retrieved support instead, it first evaluates whether the supporting evidence is sufficiently strong and consistent.

In this study, reliability refers to an operational, instance-level assessment of the quality of retrieved visual evidence, It is used to determine how the system should responds to an individual query. This meaning differs from a statistical confidence interval or a hypothesis-test significance level. Statistical confidence intervals quantify uncertainty in aggregate estimates, while significance tests evaluate evidence against a statistical null hypothesis. The retrieval-based reliability score instead supports an inference-time decision for a single input and should not be interpreted as a statistical confidence level, posterior probability, or grantee of correctness.

The reliability decision is also connected to a downstream multimodal response-generation component. When the evidence is strong and class-consistent, the predicted class can be communicated directly. When the evidence is partially supportive, the response is expressed using caution-aware language. When the evidence is weak or conflicting, the system avoids making definitive class claim and returns and abstention or fallback response. This component demonstrates how selective visual inference can be incorporated into a user-facing multimodal interface.

The quantitative evaluation is conducted in a controlled ImageNet-100 visual classification setting. This setting makes it possible to measure prediction correctness, accepted accuracy, accepted error, coverage, abstention, calibration, and high-reliability wrong predictions. The present study does not directly evaluate free-form image captioning, visual question answering, or all forms of generative hallucination in MLLMs. Instead, it investigates whether retrieval-derived evidence can help identify visual predictions that should not be confidently returned. The downstream multimodal response layer should therefore be interpreted as a proof of concept application of the selective visual decision rather than a complete multimodal hallucination-mitigation evaluation. This contribution of this work lies in the coordinated use of retrieval-based evidence, uncertainty indicators, and selective decision control within an interpretable post-hoc inference procedure. Although nearest-neighbor retrieval, entropy, evidence margins, and selective prediction are established concepts, the proposed method integrates them to evaluate whether a retrieved neighborhood provides sufficient support for a returned prediction. The experimental analysis examines the resulting reliability coverage trade-off, the differences between accepted and rejected samples, and the potential of evidence gated visual inference as a starting point for more reliable multimodal response systems.

2. Related Work

Multimodal Large Language Models (MLLMs) have become an important research direction because they combine visual perception with language-based reasoning. Earlier vision-language models focused on learning joint image-text representations for tasks such as image classification, image captioning, cross-modal retrieval, and visual question answering. Contrastive image-text pretraining methods such as CLIP demonstrated that visual concepts can be aligned with natural-language supervision and transferred effectively to downstream recognition tasks [1]. More recent multimodal systems, including LLaVA, BLIP-2, InstructBLIP, Flamingo, and MiniGPT-4, extend this capability by connecting pretrained visual encoders with large language models for open-ended visual instruction following and natural-language response generation [2, 3, 4, 5, 6].

Although these models have shown strong performance across vision-language tasks, their fluency does not always guarantee reliability. Since the final output is often expressed in natural language, users may interpret the response as grounded and trustworthy even when the underlying visual evidence is insufficient. This creates a critical challenge for applications where the system is expected not only to generate a plausible response, but also to determine whether the response is visually supported. Therefore, reliability estimation has become an important requirement for trustworthy multimodal inference.

A major limitation of current MLLMs is visual hallucination, where the generated response includes objects, attributes, relationships, or semantic claims that are not sufficiently supported by the input image. Early studies on image captioning showed that models can hallucinate objects even when they are absent from the scene [7]. More recent studies have demonstrated that hallucination remains a persistent issue in large vision-language models, especially when the image contains ambiguous visual cues, fine-grained categories, or semantically similar objects [8, 9, 10, 11].

The problem becomes more serious when hallucinated or incorrect outputs are produced with high confidence. In practical settings, an uncertain answer may be acceptable if the system clearly communicates uncertainty. However, an incorrect answer presented as confident can mislead users and reduce trust in the model. Many multimodal systems are designed to answer every input, even when the available evidence is weak or conflicting. As a result, there is a need for inference-time mechanisms that can identify unsupported or confidently incorrect visual predictions before they are converted into final user-facing responses.

Retrieval-augmented methods provide a useful direc-

tion for improving grounding and interpretability. In natural language processing, retrieval-augmented generation uses external documents or memory to support generated answers [12, 13]. Similar principles can be applied to visual inference by retrieving nearest-neighbor examples from a reference image database. In this setting, retrieved examples serve as external visual evidence that can be compared with the query image. Deep nearest-neighbor methods have shown that examples retrieved from representation space can provide interpretable support for model predictions [14]. Efficient similarity search tools such as FAISS [15] and the random projection forests based kNN search algorithm [16] make it possible to retrieve evidence from large-scale embedding databases during inference.

However, retrieval alone does not guarantee reliability. A system may retrieve visually similar examples, but the evidence may still be weak, mixed across multiple classes, or semantically ambiguous. Therefore, retrieved evidence must be evaluated before it is used to support a final prediction. This motivates the use of evidence-based reliability estimation to determine whether the retrieved examples provide reliable support for the predicted class.

Confidence calibration is closely related to trustworthy inference. Modern neural networks are often miscalibrated, meaning that their predicted confidence scores do not always match empirical correctness [17]. A model may assign high confidence to an incorrect prediction, especially under visual ambiguity, class similarity, or distribution shift. Calibration methods such as temperature scaling aim to reduce this mismatch by aligning predicted probabilities with observed accuracy. Expected Calibration Error (ECE) is commonly used to measure the gap between confidence and correctness across confidence intervals.

Uncertainty estimation methods provide additional tools for identifying unreliable predictions. Bayesian approximations, dropout-based uncertainty, deep ensembles, and related approaches estimate when a model is likely to be uncertain or wrong [18, 19, 20]. While these methods are valuable, they may require architectural modifications, additional training, or multiple forward passes. For large multimodal systems, such requirements may be computationally expensive. In contrast, retrieval-based reliability estimation can be implemented as a post-hoc inference layer that evaluates prediction trustworthiness without retraining the underlying multimodal model.

Selective prediction allows a model to abstain when confidence or evidence is insufficient [21]. Instead of forcing a prediction for every input, selective systems trade coverage for reliability. This principle is important for trustworthy artificial intelligence because abstaining from uncertain cases is often preferable to producing confident errors. Existing selective classification methods typically rely on confidence thresholds or uncertainty estimates. However, in multimodal visual inference, decision making can be strengthened by incorporating external evidence consistency. A reliability-aware decision gate can evaluate whether retrieved examples strongly support the predicted class, whether competing classes receive similar support, and whether the evidence distribution is uncertain.

The present study builds on prior work in visual representation learning, nearest-neighbor retrieval, confidence calibration, uncertainty estimation, and selective prediction. The individual components used in the proposed method are established techniques. Nearest-neighbor retrieval provides interpretable reference examples, similarity scores measure visual proximity, entropy and evidence margins characterize neighborhood uncertainty, and selective prediction allows a system to abstain when the available evidence is insufficient.

The contribution of this work lies in coordinating these components within an interpretable post-hoc decision procedure. Unlike an always-answering retrieval classifier, the proposed method evaluated the strength and class consistency of the retrieved neighborhood before determining whether a prediction should be returned. The retrieved examples are therefore used not only to generate a class prediction, but also to control whether the prediction is accepted, communicated cautiously, or rejected through abstention or fallback.

The proposed method differs from conventional confidence-thresholding approaches by using several retrieval-derived evidence signals rather than relying on a single model probability. These signals include top and mean similarity, class-support agreement, evidence margin, entropy, and aggregate reliability. However, because all of these indicators are derived from the same retrieved neighborhood, they should not be interpreted as independent evidence sources. Their combined use provides a structured assessment of retrieval strength and consistency, but it does not independently guarantee that the retrieved neighborhood is semantically correct.

Direct numerical comparison with generative MLLM hallucination-mitigation methods is not included because those methods operate on free form natural language generation and are commonly evaluated using image captioning, visual question answering or multimodal hallucination benchmarks. The present study instead evaluates closed set visual classification and selective acceptance on ImageNet-100. Because the tasks, output spaces, and evaluation measures differ, a direct numerical comparison would not represent an equivalent experimental setting. The baseline used in this work therefore isolates the effect of the proposed decision gate by comparing the same retrieval based classifier with and without selective reliability control.

3. Methodology

This section presents the retrieval-augmented reliability-aware selective inference method used to evaluate whether a visual prediction is sufficiently supported before it is returned. The method separates prediction generation from prediction acceptance. A class prediction is first produced from retrieved visual evidence, but it is returned only when the retrieved neighborhood satisfies the required evidence-strength and consistency conditions. Predictions with partially supportive evidence are communicated cautiously, while predictions with weak or conflicting evidence are rejected through abstention or fallback.

The methodology is organized into three parts. First, the data source and preprocessing procedure are described, including the ImageNet-100 reference database, image transformations, feature extraction, and embedding normalization. Second, the retrieval and selective-inference proce-

dure is presented, including FAISS-based nearest-neighbor search, class-support aggregation, reliability-signal computation, and decision control. Third, the evaluation protocol defines accepted accuracy, accepted error rate, coverage, abstention rate, Expected Calibration Error, and high-reliability wrong predictions.

The method is evaluated as a controlled visual-classification procedure. The downstream multimodal response component demonstrates how the resulting decision may control the wording of a user-facing response, but the quantitative evaluation is based on visual predictions and selective acceptance rather than free-form MLLM hallucination.

3.1. Data Source and Preprocessing Pipeline

3.1.1. Dataset and Evidence Database Split

The experiments use ImageNet-100 as a controlled visual-classification dataset. Approximately 130,000 images from the ImageNet-100 training split are used to construct the external visual evidence database, while 5,000 images from the validation split are used for quantitative evaluation.

The reference database contains the visual embeddings, class labels, and file paths of the training images. During evaluation, each validation image is treated as a query and compared with the reference database to retrieve visually similar examples. The known validation labels make it possible to determine whether the retrieval-based prediction is correct and whether the selective decision mechanism appropriately accepts or rejects the prediction.

ImageNet-100 provides a controlled setting in which the reference classes and ground-truth labels are available. However, it represents a closed-set classification task and does not capture the complete open-world complexity of multimodal systems. The experiment should therefore be interpreted as a proof-of-concept evaluation of retrieval-based selective inference.

3.1.2. Image Preprocessing

Each image is converted to RGB format and processed using the standard transformation pipeline associated with the pretrained ResNet-50 encoder. The image is resized, converted to a tensor, and normalized using ImageNet statistics. The same preprocessing procedure is applied to both the reference images and the validation queries.

Using an identical transformation pipeline is necessary because the method compares query embeddings directly with the stored reference embeddings. Differences in preprocessing could alter the resulting feature representations and make the similarity scores inconsistent.

3.1.3. Visual Feature Extraction and Normalization

A pretrained ResNet-50 model is used as the visual feature extractor. The final classification layer is removed, and the output of the penultimate layer is used as a 2,048-dimensional visual representation. For an input image x, the encoder produces the embedding:

$$z = f_\theta(x), \qquad z \in \mathbb{R}^{2048}, \tag{1}$$

where f_θ denotes the pretrained ResNet-50 feature extractor and z is the extracted visual embedding.

The embedding is then L2-normalized:

$$\tilde{z} = \frac{z}{\|z\|_2 + \epsilon}, \tag{2}$$

where ϵ is a small constant used for numerical stability. Normalization ensures that retrieval depends primarily on the direction of the embedding rather than its magnitude. Because both query and reference embeddings are normalized, the similarity search is equivalent to cosine-similarity-based retrieval.

3.2. Proposed Reliability-Aware Selective Inference Method

3.2.1. Framework Overview

Given an input image, the proposed method first converts the image into a normalized visual embedding using the pretrained ResNet-50 encoder. The query embedding is then compared with the external evidence database using FAISS-based nearest-neighbor retrieval. The top-k retrieved examples are treated as external visual evidence for estimating the most strongly supported class.

The method does not directly accept the class associated with the nearest retrieved image. Instead, the retrieved neighborhood is evaluated according to evidence strength and evidence consistency. Evidence strength describes how closely the retrieved examples match the query image, while evidence consistency describes whether the retrieved neighbors support the same class or are distributed across competing classes.

The complete selective-inference procedure contains four stages: visual encoding, external evidence retrieval, reliability-signal computation, and decision-controlled output generation. The reliability module evaluates top similarity, mean similarity, primary and secondary class support, evidence margin, entropy-based uncertainty, and aggregate reliability.

These indicators are passed to a decision gate that assigns the query image to one of three states: Accept, Answer with Caution, or Abstain/Fallback. Strong, class-consistent, low-uncertainty evidence permits acceptance; partially supportive evidence produces a cautious response; and weak or conflicting evidence triggers abstention or fallback.

As shown in Figure 1, the query image is encoded into a normalized visual embedding, compared with the external evidence database, evaluated using retrieval-derived reliability signals, and passed through the selective decision gate.

Algorithm 1 summarizes the complete inference procedure.

The algorithm emphasizes that the nearest retrieved class is not accepted automatically. The final decision depends on evidence strength, class-support consistency, uncertainty, and aggregate reliability.

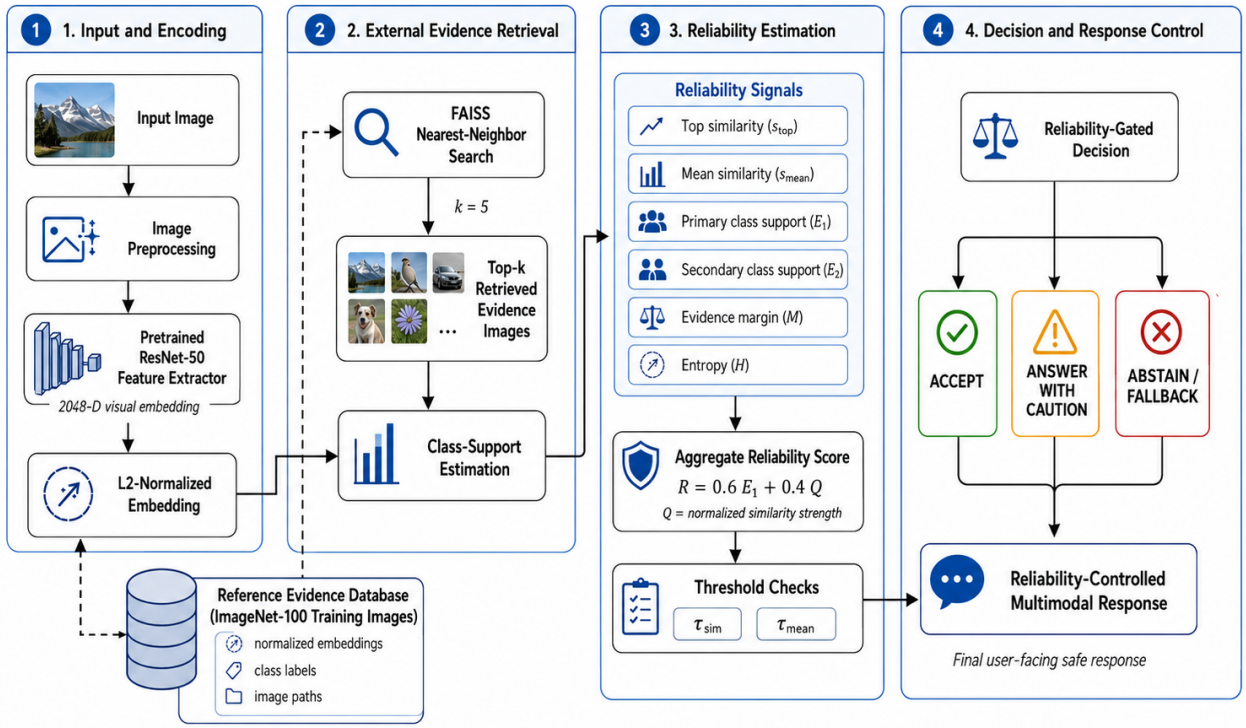

Figure 1: Overall architecture of the proposed retrieval-augmented reliability-aware selective inference method.

Algorithm 1: Retrieval-Augmented Reliability-Aware Selective Inference

Input: Query image x_q, reference database $\mathcal{R}$, visual encoder f_θ, number of neighbors k, and thresholds τ_{sim} and τ_{mean}.
Output: Decision $D(x_q)$ and reliability-controlled output.

Preprocess the query image x_q
Extract $z_q \leftarrow f_\theta(x_q)$
Normalize $\tilde{z}_q \leftarrow z_q/(\|z_q\|_2 + \epsilon)$
Retrieve $E_k(x_q) \leftarrow \{(y_j, s_j, p_j)\}_{j=1}^{k}$ using FAISS
Compute $w_j \leftarrow \exp(\alpha s_j)/\sum_{m=1}^{k} \exp(\alpha s_m)$
Compute $S(c) \leftarrow \sum_{j \in I_c} w_j$ and $\hat{y} \leftarrow \arg\max_c S(c)$
Compute s_{top}, s_{mean}, E_1, E_2, M, H, and R
Compute GateFail$(x_q) \leftarrow (s_{\text{top}} < \tau_{\text{sim}}) \vee (s_{\text{mean}} < \tau_{\text{mean}})$
if $\neg$GateFail(x_q) *and* $R \geq 0.85$ *and* $H \leq 0.30$ *and* $M \geq 0.50$ **then**
    $D(x_q) \leftarrow$ ACCEPT
else if $R \geq 0.70$ *and* $M \geq 0.25$ **then**
    $D(x_q) \leftarrow$ ANSWER WITH CAUTION
else
    $D(x_q) \leftarrow$ ABSTAIN/FALLBACK
Generate the final output according to $D(x_q)$

3.2.2. Operational Definition of Reliability

In this study, reliability refers to an instance-level operational assessment of the strength and internal consistency of the retrieved visual evidence. It is constructed from retrieval similarity, neighborhood class support, evidence margin, and entropy, and is used to determine whether an individual prediction should be accepted, communicated cautiously, or rejected.

Reliability is not used as a statistical confidence interval, hypothesis-test significance level, or posterior probability of correctness. A statistical confidence interval quantifies uncertainty in an estimated population quantity, while a significance test evaluates evidence against a specified null hypothesis. In contrast, the retrieval-based reliability score is used during inference to control the response to a single query image.

These concepts serve different purposes. Statistical confidence intervals and significance tests describe uncertainty in aggregate experimental conclusions, whereas the proposed reliability score supports a per-instance decision. Therefore, a high reliability score indicates that the retrieved neighborhood is strong and internally consistent under the proposed evidence criteria, but it does not guarantee that the prediction is correct.

The reliability indicators are not independent because they are derived from the same top-k retrieved neighborhood. The method can identify retrieval results that are weak or internally conflicting, but it cannot independently verify whether a highly consistent retrieved neighborhood is semantically correct. A systematic retrieval failure may therefore cause several reliability indicators to appear strong simultaneously. The proposed method should consequently be interpreted as a retrieval-quality-aware selective procedure rather than an independent correctness verifier.

3.2.3. Reference Evidence Database Construction

The ImageNet-100 training set is used to construct the external visual evidence database. Each training image is passed through the same preprocessing and ResNet-50 feature-extraction pipeline. For every reference image, the system stores the normalized embedding, class label, and image path. The reference evidence database is represented as:

$$\mathcal{R} = \{(\tilde{z}_i, y_i, p_i)\}_{i=1}^{N}, \tag{3}$$

where $\tilde{z}_i$ is the normalized embedding of the i-th reference image, y_i is its class label, p_i is its image path, and N is the total number of reference images.

The reference database provides external visual evidence during inference. Instead of evaluating a query image in isolation, the method compares it with labeled

reference examples and examines whether the retrieved neighborhood supports a consistent class prediction. The effectiveness of this process depends on the quality, diversity, and class coverage of the reference database. If representative examples are missing or insufficient, the retrieved evidence may be weak or misleading.

3.2.4. FAISS-Based Evidence Retrieval

To perform efficient nearest-neighbor search over the reference database, the normalized embeddings are indexed using FAISS. For a query embedding $\tilde{z}_q$, the index retrieves the top-k most similar reference examples:

$$E_k(x_q) = \{(y_j, s_j, p_j)\}_{j=1}^{k}, \tag{4}$$

where y_j is the class label of the j-th retrieved image, s_j is its similarity score with the query image, and p_j is the corresponding image path.

In the implemented system, $k = 5$ is used for reliability computation and user-facing evidence display. This value provides a small, interpretable neighborhood while allowing agreement and disagreement among multiple retrieved classes to be measured. Using one neighbor would prevent meaningful calculation of class-support consistency, evidence margin, and entropy, whereas a substantially larger neighborhood could introduce weaker visual matches and reduce the clarity of the displayed evidence. The selected value is an empirical operating choice rather than a universally optimal setting.

The retrieved images contribute to the retrieval-based class prediction and provide interpretable examples through which the supporting or challenging evidence can be inspected.

3.2.5. Retrieval-Based Class-Support Estimation

After the top-k images are retrieved, the system computes class-level evidence support. Because the nearest neighbors may belong to different classes, the prediction is not based on only the single nearest image. Instead, the method aggregates the retrieved evidence by class.

A similarity-weighted softmax function is applied to the retrieved similarity scores:

$$w_j = \frac{\exp(\alpha s_j)}{\sum_{m=1}^{k} \exp(\alpha s_m)}, \tag{5}$$

where w_j is the normalized evidence weight of the j-th retrieved image and α controls the sharpness of the weighting distribution. In the implemented experiment, $\alpha = 20$ assigns greater influence to the strongest retrieved matches while retaining contributions from all top-k neighbors. This value is an empirical operating choice rather than a universally optimal setting.

For each class c appearing in the retrieved set, the class-support score is:

$$S(c) = \sum_{j \in I_c} w_j, \tag{6}$$

where I_c is the set of retrieved neighbors assigned to class c. The predicted class is:

$$\hat{y} = \arg\max_{c} S(c). \tag{7}$$

This aggregation procedure makes the prediction depend on the distribution of support across the retrieved neighborhood rather than on a single nearest example. When the strongest retrieved images belong to the same class, the corresponding support becomes dominant. When the evidence is divided among competing classes, the support is less concentrated and the prediction is more ambiguous. However, class agreement alone does not guarantee correctness because a consistently retrieved neighborhood may still be semantically incorrect.

3.2.6. Reliability Signal Computation

The reliability module computes multiple indicators from the retrieved neighborhood to characterize evidence strength, class consistency, and ambiguity. These indicators are not statistically independent because they are derived from the same top-k examples. Their purpose is to provide complementary views of retrieval quality rather than independent guarantees of correctness.

The top similarity score is:

$$s_{\text{top}} = s_1. \tag{8}$$

It represents the similarity of the strongest retrieved image.

The mean similarity score is:

$$s_{\text{mean}} = \frac{1}{k} \sum_{j=1}^{k} s_j. \tag{9}$$

It measures the overall strength of the retrieved evidence set and prevents the system from relying only on a single strong neighbor.

The primary evidence support is:

$$E_1 = S(\hat{y}), \tag{10}$$

and the secondary evidence support is:

$$E_2 = \max_{c \neq \hat{y}} S(c). \tag{11}$$

The evidence margin is computed similarly to [22]:

$$M = E_1 - E_2. \tag{12}$$

A larger margin indicates clearer separation from competing classes.

The similarity gap is:

$$G = s_1 - s_k. \tag{13}$$

Entropy is used to estimate uncertainty in the class-support distribution:

$$H = -\sum_{c} S(c) \log\big(S(c) + \epsilon\big). \tag{14}$$

Low entropy indicates concentrated evidence, whereas high entropy indicates ambiguity or conflict among retrieved classes.

The indicators should be interpreted jointly. Top and mean similarity describe how strongly the query matches the reference database, while class support, margin, and entropy describe how consistently the retrieved neighborhood favors one class. Strong similarity with conflicting class evidence may indicate ambiguity, whereas concentrated class support with weak similarity may indicate insufficient reference coverage.

3.2.7. Reliability Score and Decision Gate

The aggregate reliability score combines primary class support and normalized similarity strength:

$$R = 0.6E_1 + 0.4Q, \tag{15}$$

where:

$$Q = \min\left(\frac{s_{\text{mean}}}{\tau_{\text{mean}}}, 1.0\right). \tag{16}$$

The weight of 0.6 assigned to class support reflects the design assumption that agreement among retrieved labels should contribute more strongly to acceptance than similarity alone. The remaining weight of 0.4 preserves the influence of neighborhood-level visual similarity. These weights were selected empirically and are not presented as statistically optimal or universally applicable.

The implemented experiment uses:

$$\begin{aligned} \tau_{\text{sim}} &= 0.8416, \\ \tau_{\text{mean}} &= 0.8285. \end{aligned} \tag{17}$$

These values correspond to lower-tail operating points derived from the observed retrieval-strength distributions. They are dataset-, encoder-, and reference-database-dependent rather than universal constants. Because an independent calibration set and complete sensitivity study were not retained, optimality or insensitivity to parameter variation is not claimed.

The evidence gate is:

$$\text{GateFail}(x) = (s_{\text{top}} < \tau_{\text{sim}}) \vee (s_{\text{mean}} < \tau_{\text{mean}}). \tag{18}$$

The final decision is:

$$D(x) = \begin{cases} A, & \neg\text{GateFail}(x),\ R \geq 0.85, \\ & H \leq 0.30,\ M \geq 0.50, \\ C, & R \geq 0.70,\ M \geq 0.25, \\ F, & \text{otherwise}, \end{cases} \tag{19}$$

where A, C, and F denote Accept, Answer with Caution, and Abstain/Fallback, respectively.

Because all decision variables originate from the same retrieved neighborhood, the gate can detect weak or internally conflicting evidence but cannot independently detect every retrieval failure. A highly consistent yet incorrect neighborhood may still produce favorable reliability indicators. The gate should therefore be understood as retrieval-quality-aware selective control rather than a guarantee of semantic correctness.

3.2.8. Reliability-Controlled Output Generation

After the decision gate determines the output state, the decision controls the final user-facing response. In the proof-of-concept pipeline, LLaVA first generates a short visual description without being directly conditioned on the retrieved class label. This separation reduces the possibility that the retrieved prediction directly biases the initial language-model description.

For Accept, the system reports the predicted class and indicates that the retrieved evidence is strong and class-consistent. For Answer with Caution, the class is presented as a possible interpretation using caution-aware language. For Abstain/Fallback, the system avoids a definitive class claim and states that the retrieval evidence is weak, ambiguous, or conflicting.

The response-generation layer is evaluated qualitatively rather than through a dedicated MLLM hallucination benchmark. The quantitative results are based on retrieval-based visual classification and selective acceptance.

For arbitrary user-uploaded images, the interface displays the predicted class, reliability score, top and mean similarity, evidence margin, entropy, decision state, retrieved examples, and controlled response. Because arbitrary uploads do not have known ground-truth labels, the demonstration reports instance-level evidence and system behavior rather than classification accuracy.

3.3. Evaluation Protocol

The evaluation procedure applies retrieval-based prediction and selective decision control to 5,000 ImageNet-100 validation images. For each query, the system retrieves the top-k reference examples, aggregates class support, computes the reliability indicators, and assigns one of the three decision states.

For quantitative analysis, outputs in the Accept and Answer with Caution states are treated as returned predictions, while Abstain/Fallback outputs are treated as rejected predictions. Accepted accuracy and accepted error are therefore calculated over all returned outputs. The caution state changes the wording of the response but does not form a separate accuracy denominator.

The evaluation uses baseline accuracy, accepted accuracy, coverage, abstention rate, accepted error rate (selective risk), Expected Calibration Error (ECE), and high-reliability wrong predictions. Baseline accuracy is:

$$\text{Acc}_{\text{base}} = \frac{N_{\text{correct}}}{N}. \tag{20}$$

Accepted accuracy is:

$$\text{Acc}_{\text{accepted}} = \frac{N_{\text{accepted correct}}}{N_{\text{accepted}}}. \tag{21}$$

Coverage and abstention are:

$$\text{Coverage} = \frac{N_{\text{accepted}}}{N}, \qquad \text{Abstention} = \frac{N_{\text{abstained}}}{N}. \tag{22}$$

Accepted error rate is:

$$\text{Risk}_{\text{accepted}} = \frac{N_{\text{accepted wrong}}}{N_{\text{accepted}}} = 1 - \text{Acc}_{\text{accepted}}. \tag{23}$$

For the always-answering baseline:

$$\text{Risk}_{\text{base}} = \frac{N_{\text{wrong}}}{N} = 1 - \text{Acc}_{\text{base}}. \tag{24}$$

Accepted error must be interpreted together with coverage because a selective method can improve returned-output accuracy by rejecting uncertain samples. This metric measures selective classification error and is not a complete measure of free-form generative hallucination.

ECE is computed over $B = 10$ bins:

$$\text{ECE} = \sum_{b=1}^{B} \frac{|B_b|}{n} \left|\text{acc}(B_b) - \text{conf}(B_b)\right|, \tag{25}$$

where B_b is the set of samples in bin b, $\mathrm{acc}(B_b)$ is the empirical accuracy of the bin, $\mathrm{conf}(B_b)$ is its average reliability score, and n is the total number of samples.

High-reliability wrong predictions are incorrect predictions with reliability scores greater than or equal to 0.85. They form a narrower subset than the total number of accepted wrong predictions.

4. Experiments and Results

This section evaluates the proposed retrieval-augmented reliability-aware inference framework on the ImageNet-100 validation set and selected user-uploaded qualitative examples. The objective is to determine whether external visual evidence and reliability-aware decision gating can reduce overconfident wrong predictions while maintaining useful prediction coverage. The baseline retrieval system always returns a prediction for every validation image. In contrast, the proposed framework accepts a prediction only when the retrieved evidence is sufficiently strong, consistent, and reliable.

The evaluation focuses on three questions. First, does the proposed framework improve the accuracy of predictions that are accepted by the system? Second, does the reliability gate reduce hallucination-like accepted wrong answers and confident wrong cases? Third, do qualitative examples show that the framework behaves appropriately under reliable, ambiguous, weak-evidence, and out-of-coverage conditions?

4.1. Experimental Setup

The experiments were conducted using the ImageNet-100 dataset. The reference evidence database was constructed from approximately 130,000 ImageNet-100 training images, and quantitative validation was performed on 5,000 validation images. Each image was passed through a pretrained ResNet-50 feature extractor with the final classification layer removed, producing a 2048-dimensional visual embedding. All embeddings were L2-normalized before indexing and retrieval.

FAISS was used to perform efficient nearest-neighbor search over the reference evidence database. For each validation image, the top-k nearest reference examples were retrieved with $k = 5$. The retrieved similarity scores were converted into class-support weights using a softmax weighting function with $\alpha = 20$. The predicted class was selected based on the highest aggregated class support among the retrieved evidence examples.

The reliability gate used two validation-derived similarity thresholds: $\tau_{\mathrm{sim}} = 0.8416$ for top similarity and $\tau_{\mathrm{mean}} = 0.8285$ for mean similarity. A prediction was accepted only when the retrieved evidence was sufficiently strong and class-consistent. Otherwise, the system either abstained or triggered fallback behavior. The baseline retrieval system was evaluated as an always-answering system, while the proposed framework was evaluated based on accepted predictions.

4.2. Evaluation Metrics

The evaluation uses baseline accuracy, accepted accuracy, coverage, abstention rate, hallucination-like accepted wrong-answer rate, Expected Calibration Error (ECE), and confident wrong cases. Baseline accuracy measures the correctness of the retrieval prediction when the system answers every validation image. Accepted accuracy measures correctness only among predictions accepted by the reliability gate. Therefore, baseline accuracy and accepted accuracy are calculated on different output sets.

For the baseline retrieval system, accuracy is computed as the number of correct predictions divided by the total number of validation samples:

$$Acc_{\mathrm{base}} = \frac{N_{\mathrm{correct}}}{N}. \tag{26}$$

Here, N_{correct} is the number of correct predictions and N is the total number of validation samples.

For the proposed reliability-aware framework, accepted accuracy is computed only over accepted predictions:

$$Acc_{\mathrm{accepted}} = \frac{N_{\mathrm{acc\_correct}}}{N_{\mathrm{acc}}}. \tag{27}$$

Here, $N_{\mathrm{acc\_correct}}$ is the number of accepted predictions that are correct and N_{acc} is the total number of accepted predictions. Coverage is the proportion of validation samples accepted by the gate:

$$Coverage = \frac{N_{\mathrm{acc}}}{N}, \tag{28}$$

and abstention rate is the proportion of validation samples rejected by the gate:

$$Abstention = \frac{N_{\mathrm{abstained}}}{N}. \tag{29}$$

In this paper, a hallucination-like visual error is defined as an incorrect visual prediction that is accepted by the system and would therefore be returned as a user-facing answer. This definition does not measure all possible forms of generative hallucination. Instead, it focuses on the reliability problem studied in this work: accepted wrong visual predictions that may appear trustworthy to the user. For the always-answering baseline, every prediction is accepted, so the hallucination-like error rate is equivalent to the ordinary prediction error rate:

$$HL_{\mathrm{base}} = \frac{N_{\mathrm{wrong}}}{N} = 1 - Acc_{\mathrm{base}}. \tag{30}$$

For the proposed framework, the hallucination-like accepted wrong-answer rate is calculated only over predictions accepted by the reliability gate:

$$HL_{\mathrm{accepted}} = \frac{N_{\mathrm{acc\_wrong}}}{N_{\mathrm{acc}}} = 1 - Acc_{\mathrm{accepted}}. \tag{31}$$

Thus, the hallucination-like error rate is related to prediction error, but it is interpreted differently for the selective framework because the system does not answer every sample. For this reason, hallucination-like error must be reported together with coverage and abstention rate. ECE is computed using the reliability score as the confidence estimate over 10 confidence bins. Confident wrong cases are defined as incorrect predictions with reliability greater than or equal to 0.85.

Table 1: Overall performance comparison between the baseline retrieval system and the proposed reliability-aware framework.

System	Acc. (%)	Accepted Error (%)	Cov. (%)	Abst. (%)	ECE (%)	Conf. Wrong
Baseline Retrieval	85.84	14.16	100.00	0.00	7.40	263
Proposed Reliability-Aware Framework	88.88	11.12	89.04	10.96	5.87	226

4.3. Overall Quantitative Results

Table 1 compares the baseline retrieval system with the proposed reliability-aware framework. In the table, the row labeled "Proposed Reliability-Aware Framework" refers to the complete framework proposed in this paper, including evidence retrieval, reliability estimation, and decision-gated response control. The baseline retrieval system uses the same retrieval-based prediction mechanism but does not apply selective reliability gating; therefore, it returns a prediction for every validation image.

The baseline retrieval system achieved an overall accuracy of 85.84%, meaning that 14.16% of the always-returned predictions were incorrect. Since the baseline accepts every prediction, its hallucination-like error rate is the same as its prediction error rate. After applying the reliability-aware evidence gate, the accepted prediction accuracy improved to 88.88%. This shows that the proposed gate filters out a portion of unreliable predictions before they are returned as final answers.

The proposed framework accepted 89.04% of the validation samples and abstained on 10.96% of uncertain or weak-evidence cases. Although this reduces coverage, it improves the trustworthiness of the predictions actually returned to the user. The hallucination-like accepted wrong-answer rate decreased from 14.16% to 11.12%, corresponding to an absolute reduction of 3.04 percentage points and a relative reduction of 21.48%. This result supports the main objective of the framework: reducing confidently accepted wrong visual outputs without retraining the underlying visual encoder or multimodal model.

Here, accepted Error denotes the proportion of incorrect predictions among the outputs returned by each system. For the baseline retrieval system, this value is equal to the ordinary prediction error rate because the baseline always returns a prediction. For the proposed framework, this value is computed only over predictions accepted by the reliability gate. Therefore, HL Error should be interpreted together with coverage and abstention rate rather than as a standalone accuracy measure.

4.4. Calibration and Reliability Analysis

The reliability-aware gate also improved calibration. The baseline retrieval system produced an ECE of 7.40%, while the proposed reliability-aware framework reduced ECE to 5.87%. This reduction indicates that the reliability score better reflects empirical correctness after unreliable samples are filtered out. In practical terms, the system becomes less likely to present unsupported predictions as reliable outputs.

The number of confident wrong cases also decreased after applying the reliability gate. The baseline system produced 263 confident wrong predictions, while the proposed framework reduced this number to 226. This reduction is important because high-confidence wrong predictions are among the most harmful visual errors in user-facing multimodal systems. By rejecting weak or conflicting evidence cases, the proposed framework reduces the risk of misleading users with confident but incorrect outputs.

4.5. Evidence Quality Analysis

Table 2 summarizes the difference between accepted samples and abstained/fallback samples. Accepted samples achieved an accuracy of 88.88%, while abstained/fallback samples had a much lower accuracy of 61.13%. This confirms that the reliability gate is not rejecting samples randomly; instead, it tends to reject cases that are genuinely harder or less reliable.

Accepted samples also showed stronger evidence signals. Their average reliability score was 0.9560, compared with 0.8137 for abstained/fallback samples. The accepted group had higher average top similarity, higher mean similarity, larger evidence margin, and lower entropy. Specifically, accepted samples had an average margin of 0.8666 and entropy of 0.1563, while abstained/fallback samples had a lower margin of 0.5380 and a higher entropy of 0.6308. These results show that accepted predictions are supported by stronger and more class-consistent retrieved evidence, while abstained/fallback samples contain weaker or more ambiguous evidence.

4.6. Qualitative Case Analysis

In addition to the quantitative evaluation, qualitative case analysis was conducted to examine how the proposed framework behaves at the individual-image level. This analysis is important because the objective of the proposed method is not only to improve aggregate accuracy, but also to control the final response based on the strength and consistency of retrieved evidence. A reliable multimodal visual system should accept predictions when evidence is strong, but should avoid confident answers when the retrieved evidence is weak, ambiguous, conflicting, or outside the reference class distribution.

Table 3 summarizes four representative qualitative examples. These examples include one high-confidence accepted example and three abstain/fallback examples. The accepted case demonstrates that the proposed framework does not unnecessarily reject reliable predictions. The fallback cases demonstrate how the system handles ambiguous visual evidence, degraded image quality, and out-of-coverage inputs.

Table 2: Evidence quality comparison between accepted and abstained/fallback samples.

Group	Samples	Acc.	Rel.	s_{top}	s_{mean}	Margin	Entropy
Accepted	4452	0.8888	0.9560	0.9110	0.9000	0.8666	0.1563
Abstained/Fallback	548	0.6113	0.8137	0.8168	0.8028	0.5380	0.6308

Table 3: Qualitative case study examples showing reliability-aware decision outcomes.

Case	Input	Pred.	Decision	Rel.	Reason
1	Shark	Shark	ACCEPT	1.0000	Strong same-class evidence
2	Birds	Kite	FALLBACK	0.5931	Mixed bird-class evidence
3	Blurred car	Turtle	FALLBACK	0.4316	Weak unrelated retrieval
4	Airplane	Crane	FALLBACK	0.5505	Outside reference coverage

Case 1: Accepted prediction–great white shark. The first qualitative case, shown in Figure 2 and Figure 3, demonstrates a high-confidence accepted prediction. The input image contains a great white shark, and all top-five retrieved evidence examples belong to the same class. The reliability score is 1.0000, the top similarity is 0.9511, the mean similarity is 0.9361, the evidence margin is 1.0000, and the entropy is 0.0000. Since the retrieved examples are visually close, class-consistent, and free from competing class evidence, the decision gate accepts the prediction.

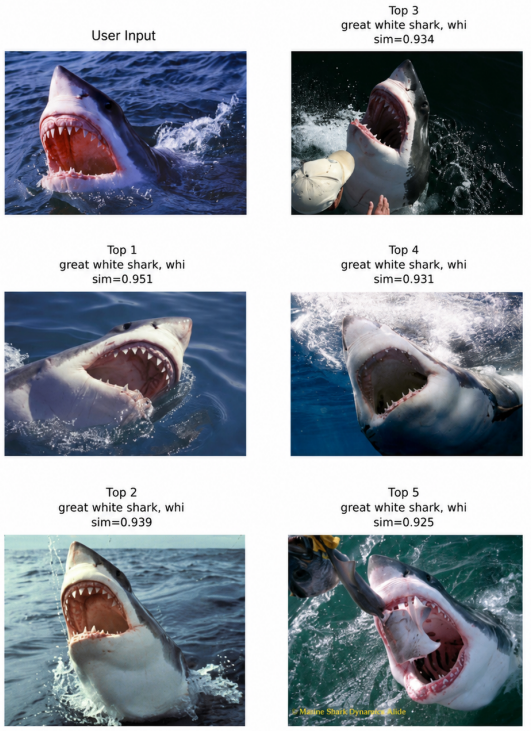

Figure 2: Accepted qualitative case showing the input image and top-5 retrieved evidence for the great white shark example.

Case 1: Accepted Prediction — Great White Shark

Reliability Metrics

Decision:	ACCEPT
Predicted class:	Great white shark
Reliability:	1.0000
Top similarity:	0.9511
Mean similarity:	0.9361
Margin:	1.0000
Entropy:	0.0000

Reason for Decision

All retrieved evidence belongs to the same class and shows high visual similarity. The reliability gate thresholds are satisfied, so the prediction is accepted.

Final Controlled Response

The image is of a great white shark swimming in the ocean.

Figure 3: Reliability summary and final controlled response for the accepted great white shark example.

Case 2: Abstain/fallback–birds in flight. The second case, shown in Figure 4 and Figure 5, contains a group of blurred birds flying in the sky. The retrieval system predicts the class as kite. However, the top retrieved evidence is distributed across multiple bird-related classes, including goose, crane, kite, and magpie. The reliability score is 0.5931, the top similarity is 0.7854, the mean similarity is 0.7741, the evidence margin is 0.1172, and the entropy is 1.3480. The low margin and high entropy indicate class conflict among the retrieved examples, so the system assigns the case to ABSTAIN/FALLBACK.

Figure 4: Abstain/fallback qualitative case showing the input image and top-5 retrieved evidence for the birds-in-flight example.

Case 2: Abstain/Fallback — Birds in Flight

Reliability Metrics

Decision:	ABSTAIN / FALLBACK
Predicted class:	Kite
Reliability:	0.5931
Top similarity:	0.7854
Mean similarity:	0.7741
Margin:	0.1172
Entropy:	1.3480

Reason for Decision

The retrieved evidence is mixed across multiple bird-related classes, including goose, crane, kite, magpie, and kite. The evidence margin is low and entropy is high, so the system avoids a confident prediction.

Final Controlled Response

The image appears to show birds flying in the sky, but the retrieved evidence is ambiguous. The system abstains from assigning a confident class label.

Figure 5: Reliability summary and final controlled response for the birds-in-flight fallback example.

Case 3: Abstain/fallback–blurred car. The third case, shown in Figure 6 and Figure 7, contains a heavily blurred car image. The nearest retrieved class is loggerhead turtle, which is not semantically appropriate for the uploaded image. The retrieved evidence includes loggerhead turtle, snail, tarantula, goldfish, and hummingbird. The reliability score is 0.4316, the top similarity is 0.6218, the mean similarity is 0.6162,

the evidence margin is 0.0249, and the entropy is 1.6074. These values indicate weak and conflicting evidence, so the system abstains instead of forcing an incorrect label.

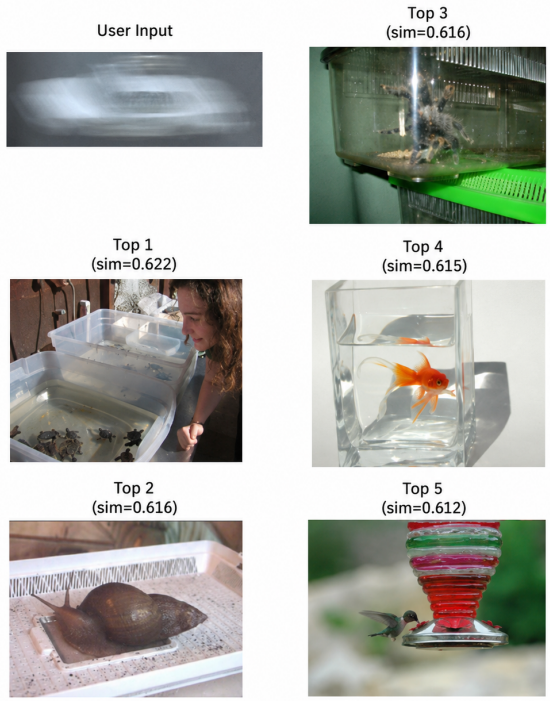

Figure 6: Abstain/fallback qualitative case showing the input image and top-5 retrieved evidence for the blurred car example.

Case 3: Abstain/Fallback — Blurred Car Image

Reliability Metrics	
Decision:	ABSTAIN / FALLBACK
Nearest class:	Loggerhead turtle
Reliability:	0.4316
Top similarity:	0.6218
Mean similarity:	0.6162
Margin:	0.0249
Entropy:	1.6074

Reason for Decision

The input image is heavily blurred, and the retrieved classes are weak or conflicting. The nearest known class is unrelated to the visual input, so the framework rejects the prediction instead of returning a confident answer.

Final Controlled Response

A blurred image of a car is presented, but the visual evidence is insufficient to confidently identify the exact class using the reference evidence base.

Figure 7: Reliability summary and final controlled response for the blurred car fallback example.

Case 4: Abstain/fallback–airplane. The fourth case, shown in Figure 8 and Figure 9, contains a commercial airplane. The retrieval system predicts the class as crane because airplane is not supported by the ImageNet-100 reference subset used in this experiment. The top retrieved classes include crane, tiger shark, kite, goose, and hammerhead shark. The reliability score is 0.5505, the top similarity is 0.7571, the mean similarity is 0.7296, the evidence margin is 0.1257, and the entropy is 1.5571. Although the visual-language response can describe the image as an airplane, the retrieval evidence does not support a reliable ImageNet-100 class prediction, so the reliability gate assigns the case to ABSTAIN/FALLBACK.

Figure 8: Abstain/fallback qualitative case showing the input image and top-5 retrieved evidence for the airplane example.

Case 4: Abstain/Fallback — Airplane Image

Reliability Metrics	
Decision:	ABSTAIN / FALLBACK
Predicted class:	Crane
Reliability:	0.5505
Top similarity:	0.7571
Mean similarity:	0.7296
Margin:	0.1257
Entropy:	1.5571

Reason for Decision

The input image is a commercial airplane, but the retrieved evidence is distributed across unrelated classes, including crane, tiger shark, kite, goose, and hammerhead shark. Because the evidence is inconsistent and outside the reference class coverage, the framework abstains from returning a confident class label.

Final Controlled Response

The image shows a large commercial airplane, likely a United Airlines plane, flying in the sky, but the system abstains because the retrieved class evidence is outside the supported reference coverage.

Figure 9: Reliability summary and final controlled response for the airplane fallback example.

Overall, these qualitative examples demonstrate that the proposed framework behaves in a reliability-aware manner across different input conditions. The shark example shows acceptance under strong and consistent evidence. The birds example shows abstention under class ambiguity among visually related categories. The blurred car example shows fallback under low-quality and weak-evidence conditions. The airplane example shows protection against out-of-coverage inputs. Together, these cases show that the framework reduces overconfident visual hallucinations by preventing unsupported predictions from being accepted as final answers.

4.7. Result Discussion

The experimental findings demonstrate the expected reliability–coverage trade-off of selective inference. The always-answering retrieval baseline returns a prediction for every query, including cases supported by weak, am-

biguous, or conflicting evidence. The proposed decision mechanism rejects a subset of these cases and consequently improves the correctness of the outputs that are returned. This improvement should therefore be interpreted together with the corresponding reduction in coverage rather than as an unrestricted increase in classification performance.

The evidence-characteristic analysis helps explain the behavior of the decision gate. Returned predictions generally exhibit stronger top and mean similarity, larger evidence margins, more concentrated class support, and lower entropy than rejected predictions. The underlying retrieval classifier also performs substantially worse on the rejected group, indicating that the gate tends to identify empirically more difficult samples rather than rejecting queries arbitrarily.

However, the reliability indicators are all derived from the same retrieved neighborhood and are therefore correlated. The method can identify retrieval results that are weak or internally inconsistent, but it cannot independently verify that a highly consistent neighborhood is semantically correct. Consequently, the reported improvements demonstrate the usefulness of retrieval-quality-aware selective control under the implemented ImageNet-100 setting, but they do not establish that every accepted prediction is correct.

The comparison with the always-answering baseline isolates the effect of the proposed decision gate because both systems use the same image embeddings, retrieval process, similarity weighting, and class-support prediction. The study does not claim numerical superiority over all existing calibration, uncertainty-estimation, or selective-classification methods, since those alternatives were not evaluated in the retained experiment. Similarly, direct comparison with generative MLLM hallucination-mitigation approaches is not methodologically equivalent because those methods operate on free-form language outputs and use different datasets and evaluation measures.

The qualitative examples further illustrate the intended system behavior. Strong and class-consistent retrieval evidence permits a prediction to be returned, while ambiguous, degraded, or out-of-reference evidence can lead to fallback. These examples support the use of retrieval-derived reliability as a transparent decision aid, but they do not constitute a quantitative evaluation of out-of-distribution detection or free-form multimodal hallucination.

Overall, the results should be interpreted as a controlled proof-of-concept for retrieval-augmented selective visual inference. The method provides an interpretable post-hoc mechanism for deciding whether a visual prediction should be returned, qualified, or rejected before it is incorporated into a downstream multimodal response. Broader conclusions will require independent parameter calibration, sensitivity analysis, additional selective-prediction baselines, quantitative out-of-distribution testing, and evaluation on dedicated multimodal-generation benchmarks.

5. Limitations

The present study has several limitations. First, the quantitative evaluation is conducted on ImageNet-100, which represents a controlled closed-set visual-classification setting. Although this setting allows prediction correctness, selective risk, coverage, and abstention to be measured clearly, it does not represent the full complexity of open-ended multimodal tasks such as image captioning, visual question answering, or free-form MLLM response generation. The downstream multimodal response component should therefore be interpreted as a proof-of-concept demonstration rather than a complete evaluation of hallucination mitigation.

Second, the proposed reliability indicators depend on the same top-k retrieved neighborhood. Top similarity, mean similarity, class support, evidence margin, entropy, and aggregate reliability are therefore correlated rather than independent sources of evidence. The method can identify retrieval results that are weak or internally conflicting, but it cannot independently verify whether a highly consistent retrieved neighborhood is semantically correct. A systematic retrieval failure may cause several reliability indicators to appear favorable simultaneously.

Third, the method depends on the quality, diversity, and class coverage of the external evidence database. When the target concept is absent or poorly represented, the retrieval procedure may return visually similar but semantically unrelated examples. Weak similarity or conflicting class support may trigger fallback in some such cases, but the proposed method is not a formally calibrated out-of-distribution detector. The airplane example provides only a qualitative illustration of out-of-reference behavior and should not be interpreted as a quantitative OOD evaluation.

Fourth, the number of neighbors, similarity-weighting parameter, reliability-score weights, and decision thresholds were selected empirically for the implemented configuration. These values are dependent on the dataset, encoder, and reference database and are not claimed to be universally optimal. A complete sensitivity analysis and an independently retained calibration split were not available for the present evaluation. Future studies should separate parameter calibration from final testing and evaluate performance across multiple threshold and weighting configurations.

Fifth, the experimental comparison is limited to an always-answering retrieval baseline that uses the same embeddings, retrieval mechanism, and class-support prediction without selective decision control. This comparison isolates the effect of the proposed gate, but it does not establish superiority over existing calibration, uncertainty-estimation, or selective-classification methods. Direct numerical comparison with generative MLLM hallucination-mitigation approaches is also not methodologically equivalent in the current setting because those methods operate on free-form language outputs and use different datasets and evaluation measures.

Finally, the reliability-controlled language-generation layer is evaluated primarily through qualitative examples. Although the decision state controls whether the final response is direct, cautious, or abstaining, the language model may still produce unsupported details in an open-ended setting. Broader validation will require dedicated multimodal hallucination benchmarks, stronger vision-language encoders, larger and more diverse evidence databases, quantitative OOD evaluation, additional selective-prediction baselines, and systematic parameter-sensitivity analysis.

6. Discussion and Conclusion

This study presented a retrieval-augmented reliability-aware selective inference method for visual classification. The method separates prediction generation from prediction acceptance by evaluating whether the retrieved visual neighborhood provides sufficiently strong and consistent support for a returned prediction. Instead of forcing an answer for every query, the system can return a prediction, communicate it cautiously, or invoke abstention or fallback according to the quality of the retrieved evidence.

The proposed procedure combines normalized visual embeddings, FAISS-based nearest-neighbor retrieval, similarity-weighted class support, evidence margin, entropy-based uncertainty, and an aggregate reliability score. These indicators are used operationally to control the output for an individual query. They should not be interpreted as statistical confidence intervals, significance levels, posterior probabilities, or guarantees of correctness.

The ImageNet-100 evaluation demonstrates a reliability–coverage trade-off. The accuracy of returned predictions increases from 85.84% for the always-answering retrieval baseline to 88.88% for the selective method at 89.04% coverage. The accepted error rate decreases from 14.16% to 11.12%, corresponding to an absolute reduction of 3.04 percentage points and a relative reduction of 21.48%. Expected Calibration Error decreases from 7.40% to 5.87%, while the number of high-reliability wrong predictions decreases from 263 to 226. These results indicate that rejecting a subset of weak or conflicting retrieval cases can improve the correctness of the predictions that are returned.

The evidence analysis further shows that returned predictions generally have stronger similarity, larger evidence margins, more concentrated class support, and lower entropy than rejected predictions. The qualitative examples illustrate acceptance under strong class-consistent evidence and fallback under ambiguous, degraded, or out-of-reference retrieval conditions. However, these examples do not constitute a quantitative evaluation of out-of-distribution detection or free-form multimodal hallucination.

The contribution of this work lies in coordinating established retrieval, uncertainty, and selective-prediction concepts within an interpretable post-hoc decision procedure. The current experiment does not establish superiority over all calibration, uncertainty-estimation, or selective-classification methods. It should also not be interpreted as a complete framework for reducing generative hallucination in MLLMs. Instead, it provides a controlled proof-of-concept showing how retrieval-derived evidence can influence whether a visual prediction should be returned, qualified, or rejected before being incorporated into a downstream multimodal response.

Future work should evaluate the method using independent calibration and test sets, systematic sensitivity analysis, additional selective-prediction baselines, quantitative out-of-distribution benchmarks, larger and more diverse evidence databases, and stronger vision-language encoders. Extension to visual question answering, image captioning, and dedicated multimodal hallucination benchmarks will be necessary to determine whether the proposed selective inference principles generalize to open-ended MLLM responses.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment This work is partially supported by the Office of Naval Research (ONR) Grant, MUST IV (N00014-23-1-2141), University of Massachusetts Dartmouth.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, "Learning transferable visual models from natural language supervision", "Proceedings of the 38th International Conference on Machine Learning", vol. 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, PMLR, 2021.

[2] H. Liu, C. Li, Q. Wu, Y. J. Lee, "Visual instruction tuning", "Advances in Neural Information Processing Systems", vol. 36, pp. 34892–34916, Curran Associates, Inc., 2023, doi:10.52202/075280-1516.

[3] J. Li, D. Li, S. Savarese, S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models", "Proceedings of the 40th International Conference on Machine Learning", vol. 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742, PMLR, 2023.

[4] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, S. C. H. Hoi, "InstructBLIP: Towards general-purpose vision-language models with instruction tuning", "Advances in Neural Information Processing Systems", vol. 36, pp. 49250–49267, Curran Associates, Inc., 2023, doi:10.52202/075280-2142.

[5] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. L. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Bińkowski, R. Barreira, O. Vinyals, A. Zisserman, K. Simonyan, "Flamingo: A visual language model for few-shot learning", "Advances in Neural Information Processing Systems", vol. 35, pp. 23716–23736, Curran Associates, Inc., 2022, doi:10.52202/068431-1723.

[6] D. Zhu, J. Chen, X. Shen, X. Li, M. Elhoseiny, "MiniGPT-4: Enhancing vision-language understanding with advanced large language models", "International Conference on Learning Representations", 2024.

[7] A. Rohrbach, L. A. Hendricks, K. Burns, T. Darrell, K. Saenko, "Object hallucination in image captioning", "Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing", pp. 4035–4045, Association for Computational Linguistics, Brussels, Belgium, 2018, doi:10.18653/v1/D18-1437.

[8] Y. Li, Y. Du, K. Zhou, J. Wang, X. Zhao, J.-R. Wen, "Evaluating object hallucination in large vision-language models", "Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing", pp. 292–305, Association for Computational Linguistics, Singapore, 2023, doi:10.18653/v1/2023.emnlp-main.20.

[9] Z. Bai, P. Wang, T. Xiao, T. He, Z. Han, Z. Zhang, M. Z. Shou, "Hallucination of multimodal large language models: A survey", *arXiv preprint arXiv:2404.18930*, 2024, doi:10.48550/arXiv.2404.18930.

[10] W. Huang, H. Liu, M. Guo, N. Gong, "Visual hallucinations of multi-modal large language models", "Findings of the Association for Computational Linguistics: ACL 2024", pp. 9614–9631, Association for Computational Linguistics, Bangkok, Thailand, 2024, doi:10.18653/v1/2024.findings-acl.573.

[11] S. Leng, H. Zhang, G. Chen, X. Li, S. Lu, C. Miao, L. Bing, "Mitigating object hallucinations in large vision-language models through visual contrastive decoding", "Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition", pp. 13872–13882, 2024, doi:10.1109/CVPR52733.2024.01316.

[12] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks", "Advances in Neural Information Processing Systems", vol. 33, pp. 9459–9474, Curran Associates, Inc., 2020.

[13] K. Guu, K. Lee, Z. Tung, P. Pasupat, M. Chang, "Retrieval augmented language model pre-training", "Proceedings of the 37th International Conference on Machine Learning", vol. 119 of *Proceedings of Machine Learning Research*, pp. 3929–3938, PMLR, 2020.

[14] N. Papernot, P. McDaniel, "Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning", *arXiv preprint arXiv:1803.04765*, 2018, doi:10.48550/arXiv.1803.04765.

[15] J. Johnson, M. Douze, H. Jégou, "Billion-scale similarity search with GPUs", *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2021, doi:10.1109/TBDATA.2019.2921572.

[16] D. Yan, Y. Wang, J. Wang, H. Wang, Z. Li, "K-nearest neighbor search by random projection forests", *IEEE Transactions on Big Data*, vol. 7, no. 1, pp. 147–157, 2021, doi:10.1109/TBDATA.2019.2908178.

[17] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, "On calibration of modern neural networks", "Proceedings of the 34th International Conference on Machine Learning", vol. 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, PMLR, 2017.

[18] Y. Gal, Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning", "Proceedings of the 33rd International Conference on Machine Learning", vol. 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059, PMLR, 2016.

[19] B. Lakshminarayanan, A. Pritzel, C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles", "Advances in Neural Information Processing Systems", vol. 30, pp. 6402–6413, Curran Associates, Inc., 2017.

[20] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, J. Snoek, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift", "Advances in Neural Information Processing Systems", vol. 32, pp. 13991–14002, Curran Associates, Inc., 2019.

[21] Y. Geifman, R. El-Yaniv, "Selective classification for deep neural networks", "Advances in Neural Information Processing Systems", vol. 30, pp. 4878–4887, Curran Associates, Inc., 2017.

[22] D. Yan, P. Wang, M. Linden, B. S. Knudsen, T. W. Randolph, "Statistical methods for tissue array images—algorithmic scoring and co-training", *The Annals of Applied Statistics*, vol. 6, no. 3, pp. 1280–1305, 2012, doi:10.1214/12-AOAS543.

Pratheswaran Hariharan received his Master of Science degree in Data Science from the University of Massachusetts Dartmouth, USA, in May 2026. He has worked as a Graduate Research Assistant focusing on reliability-aware artificial intelligence systems, multimodal reasoning, confidence-aware inference, and evidence-based decision mechanisms.

His research interests include multimodal large language models, visual hallucination mitigation, retrieval-augmented inference, embedding-based evidence retrieval, confidence calibration, uncertainty estimation, selective prediction, and trustworthy AI. He is particularly interested in developing post-hoc reliability layers that allow AI systems to evaluate when predictions should be accepted, softened, or rejected based on external evidence. His broader research goal is to build safer, interpretable, and reliable multimodal AI systems for real-world decision-support applications.

Haiping Xu received the B.S. degree in Electrical Engineering from Zhejiang University, China, in 1989, the M.S. degree in Computer Science from Wright State University in 1996, and the Ph.D. degree in Computer Science from the University of Illinois at Chicago in 2003. Before 1996, he worked as a software engineer at Shen-Yan Systems Technology, Inc. and Hewlett-Packard in Beijing. Since 2003, he has been with University of Massachusetts Dartmouth, where he is a Professor and Chairperson of the Computer and Information Science Department and Director of the AI and Software Engineering (AISER) Laboratory.

His expertise includes artificial intelligence and distributed software engineering. His research interests span cloud computing, software reliability, formal methods, cybersecurity, multi-agent systems, and deep learning. His work has been supported by the NSF, the U.S. Marine Corps, ONR, and the UMass President's Office. He is a senior member of the IEEE and ACM.

Donghui Yan received the B.S. and M.S. degrees in Applied Mathematics from Shanghai Jiao Tong University in 1991 and 1994, respectively, and his Ph.D. in Statistics from the University of California, Berkeley, in 2008. He is currently an Associate Professor of Mathematics and Data Science and Program Director of the Data Science program at the University of Massachusetts Dartmouth. Previously, he worked as a research scientist at Intel Research Berkeley and as a principal data scientist at Walmart Labs.

His research interests include statistics, machine learning, and data mining. His work focuses on statistical modeling, scalable learning algorithms, ensemble and randomized algorithms, and data-driven decision-making in complex systems. He has contributed to large-scale data analysis and predictive modeling, with recent work on environmental modeling, intelligent systems, and sensor-driven analytics for real-world applications.