From Weak Embeddings to Trustworthy Multimodality: Modality-Agnostic Learning with Confidence-Aware Fallback

Ajmal Abbas[1], Pratheswaran Hariharan[1], Aswin Sankar[1], Vishnu Selvaraj[1], Long Jiao[1]
[1]Department of Computer and Information Science
University of Massachusetts Dartmouth
Dartmouth, MA, USA
{aabbas1, phariharan, asankar, vselvaraj, ljiao}@umassd.edu

***Abstract*—Multimodal large language models (LLMs) enable unified understanding across diverse data modalities such as images, audio, video, and text by projecting them into a shared embedding space. However, existing approaches predominantly rely on image-centric supervision and uncalibrated softmax-based confidence estimates, leading to semantically weak embeddings, modality imbalance, and overconfident predictions that do not reflect true correctness. This limitation reduces the reliability of multimodal systems in decision-critical applications. To address these challenges, we propose a modality-agnostic multimodal learning framework that jointly integrates representation learning with confidence calibration, uncertainty estimation, and a confidence-aware fallback mechanism. The framework employs semantically grounded multimodal alignment, temperature scaling for post-hoc calibration, and entropy-based uncertainty modeling to improve confidence reliability. In contrast to prior work that treats calibration independently, our approach incorporates confidence estimation directly into the multimodal pipeline and leverages cross-modal reasoning through a fallback mechanism to validate low-confidence predictions. We evaluate the proposed framework on image and video modalities using benchmark datasets, including CIFAR-100 and an action recognition dataset. Performance is assessed using Expected Calibration Error (ECE), reliability diagrams, and prediction accuracy. Experimental results demonstrate a significant improvement in calibration quality, with ECE reduced from 0.1054 to 0.0438, alongside improved alignment between predicted confidence and empirical accuracy. Additionally, the fallback mechanism enhances robustness by reducing incorrect high-confidence predictions. These results indicate that the proposed framework not only improves predictive performance but also produces trustworthy and well-calibrated confidence estimates, enabling reliable deployment of multimodal AI systems in real-world, decision-critical scenarios.**

***Keywords*—Multi modal learning, Semantic embeddings, Contrastive representation learning, Confidence Calibration, Uncertainty Estimation, Vision language models.**

I. Introduction

Multimodal learning has improved many tasks by combining information from text, images, and video in a shared representation space. This setting supports classification, retrieval, and zero-shot inference. However, strong prediction accuracy alone is not enough for practical use. A multimodal system should also know when its prediction is reliable and when it is uncertain. This need is important because many current systems still produce overconfident predictions. In practice, softmax scores are often used as confidence values, but these scores do not always match the true likelihood of correctness. The problem becomes more serious in multimodal settings, where errors from one modality can affect the final decision. As a result, a model may appear confident even when the input is ambiguous or the prediction is wrong. Most existing multimodal methods focus on representation learning and task accuracy, but they give much less attention to reliability. In many pipelines, predictions from retrieval, classification, or reasoning modules are used directly without a clear verification step. This limits robustness and interpretability, especially when the data are noisy, incomplete, or semantically inconsistent. For real-world use, multimodal systems need not only strong features, but also calibrated confidence and a clear way to handle uncertain cases. To address this issue, we propose a confidence-aware multimodal framework for reliable inference. The framework first learns aligned visual and semantic representations with a pretrained multimodal encoder. It then improves prediction reliability through temperature scaling and entropy-based uncertainty estimation. When the model is not confident, a fallback mechanism uses cross-modal reasoning to re-check the prediction. This design supports both zero-shot inference and supervised fine-tuning while keeping the decision process more transparent. The main contribution of this work is a simple unified framework that combines representation learning, confidence calibration, uncertainty estimation, and fallback validation in one pipeline. The goal is not only to improve prediction quality, but also to make the output easier to trust in decision-critical settings.

II. Related Work

Recent work in multimodal learning has mainly focused on mapping different data types into a shared embedding space for unified semantic understanding. CLIP [1] is a key example. It showed that large-scale contrastive learning can align images and text effectively and support strong zero-shot performance. Building on this idea, ImageBind [2] extended shared representation learning to more modalities and showed that a common embedding space can support broader multimodal reasoning.

These studies established shared embedding learning as a central direction in multimodal research.

Later work expanded this contrastive framework to additional domains. AudioCLIP [3] added audio to the vision-language space and improved cross-modal alignment across sound, image, and text. VideoCLIP [4] aligned video and text with temporal modeling, which improved performance in video understanding tasks. In the 3D setting, PointCLIP [5] adapted pretrained vision-language models to point cloud data through multi-view projection. Together, these methods show that shared embedding learning can scale well across different modalities. However, they still rely mainly on similarity-based inference. They improve feature quality, but they do not directly measure whether a final prediction is reliable or uncertain.

Another issue in this line of work is modality imbalance. Many multimodal systems are still built around image-centered supervision or image-text alignment, even when they are extended to other modalities. This design often improves performance, but it can also make it harder to support balanced and symmetric reasoning across modalities. More recent methods try to reduce this limitation. UniBind [6], for example, introduces modality-agnostic embedding centers built from LLM-generated textual descriptions. This design improves semantic consistency across heterogeneous inputs and supports both zero-shot and fine-tuned settings. Even so, the main focus remains representation quality. The downstream prediction stage still lacks an explicit mechanism to decide whether a model output should be trusted.

This gap is important because real multimodal systems often work with noisy, incomplete, or semantically inconsistent inputs. In such cases, strong embeddings alone are not enough. A model may still produce a confident answer even when the input is ambiguous or the evidence across modalities is weak. As a result, better representation learning does not automatically lead to reliable inference. This motivates the need to study confidence and uncertainty together with multimodal alignment, rather than as separate problems.

Separate from multimodal representation learning, another line of work studies confidence calibration and uncertainty estimation in neural networks. Guo et al. [7] showed that modern deep models are often poorly calibrated and introduced temperature scaling as a simple post-hoc solution. Their work showed that predicted probabilities may look confident even when they do not match true correctness. Temperature scaling became widely used because it is simple, effective, and easy to apply after training. However, most calibration methods were developed for unimodal classification settings and are not tightly integrated into multimodal inference pipelines.

Ovadia et al. [8] further showed that uncertainty estimates can degrade under distribution shift, which limits their reliability in practical settings. This result is especially important for multimodal systems, where the input may vary widely across domains, data quality, and modality combinations. In vision-language systems, the problem can become even more serious, since models may generate high-confidence outputs that are semantically wrong or hallucinated. These findings suggest that confidence estimation should not be treated as a small post-processing step only. It should be connected to the full inference process.

To improve reliability at inference time, prior studies have explored selective prediction and confidence-aware decision strategies. SelectiveNet [9] allows a model to abstain from prediction when confidence is low by learning prediction and selection jointly. Deep Gamblers [10] uses a related idea by assigning probability mass to an abstention option. These methods improve caution in uncertain cases, but they are still mostly designed for single-model prediction settings. They do not fully use cross-modal evidence to verify an answer or revise a weak prediction through additional reasoning.

Recent studies on large language models also examine confidence estimation in more complex settings. Sun et al. [11] study confidence scoring for dialogue state tracking using softmax-based, token-level, verbalized, and combined signals, with additional self-probing and fine-tuning. Their results show that confidence modeling can improve reliability, but the method also adds computational overhead and model dependency. Spatharioti et al. [12] analyze how LLM-based search affects human decisions and show that faster decisions may also increase over-reliance on model outputs. This work highlights the practical importance of trust calibration in human-AI settings.

Overall, existing multimodal methods mainly improve representation learning, while calibration and uncertainty are usually studied as separate problems. Selective prediction methods add caution, but they do not fully use cross-modal signals to verify uncertain outputs. Recent LLM confidence studies provide useful ideas, but they are not designed as unified multimodal inference frameworks. Therefore, there is still a need for a simple framework that combines multimodal representation learning with calibrated confidence, uncertainty estimation, and cross-modal validation during inference.

III. Methodology

A. Multi modal Input Representation and Prepossessing

The proposed framework operates on visual inputs and their corresponding semantic representations, forming a vision-language multi modal system. Unlike general multi modal pipelines that incorporate video or temporal data, this work focuses on static visual inputs augmented with semantic reasoning to ensure reliability-aware inference. Prior to inference, modality-specific preprocessing is applied to ensure consistency and compatibility with pretrained models. For the visual modality, input images are transformed into standardized tensor representations through resizing and normalization, following the requirements of the underlying feature extraction backbone. This step reduces variability arising from heterogeneous image dimensions and stabilizes feature extraction. In parallel, auxiliary semantic representations are generated using vision-language models, which convert visual content into descriptive textual cues. These semantic signals capture high-level contextual information and are later utilized for cross-modal validation and fallback reasoning.

B. Modality-specific Feature Extraction

The modality-specific encoders are employed to extract discriminative feature representations from each modality. For the visual branch, a pretrained deep neural network is used to extract high dimensional feature embeddings:

$$\phi(x) \in \mathbb{R}^d \tag{1}$$

where $\phi(x)$ denotes the embedding of the input image x, and $d = 2048$.

These embeddings capture both structural and sematic information and serve as the primary representation for classification, similarity-based retrieval, and evidence-aware validations. In addition, sematic representations derived from vision-language models provide complementary information, enabling a richer multi modal representation.

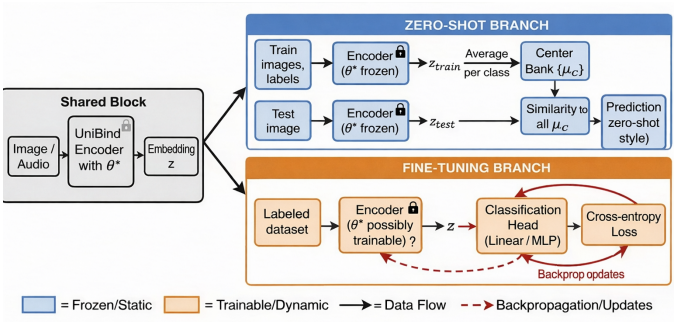

Fig. 1: Downstream inference framework leveraging shared embeddings for both zero-shot classification via center-based similarity and supervised fine-tuning through a trainable classification head.

C. Shared Embedding Learning for cross-modal alignment

To ensure cross-modal consistency, visual and sematic representations are aligned within a shared embedding space. Let $\phi(x)$ and $\psi(x)$ denote the visual and semantic embeddings, respectively. These representations are projected into a common latent space using learnable transformations:

$$z^{(v)} = g_v(\phi(x)), \quad z^{(s)} = g_s(\psi(x)) \tag{2}$$

where $g_v(\cdot)$ and $g_s(\cdot)$ are modality-specific projection functions. The projected embeddings are subsequently normalized and compared using similarity measures to capture semantic correspondence across modalities. This alignment allows cross-modal agreement to serve as an additional reliability signal, complementing the primary model predictions. To enable flexible downstream inference, the shared embeddings for both zero-shot and fine-tuning paradigms. In zero-shot branch, class-level representations are formed using a center bank and predictions are obtained via similarity matching. In parallel, the fine-tuning branch employs a trainable classification head optimized using labeled data. This dual-branch design allows the framework to operate in both data-scarce and supervised settings.

D. Base Prediction and Softmax-Based Confidence Estimation

The base prediction is generated using a fully connected linear classification layer operating on the visual embeddings extracted from the pretrained ResNet50 backbone. This layer maps the 2048-dimensional feature representation to a k-dimensional logit vector, where$k$ is the number of target classes. Let $z \in \mathbb{R}^K$ denote the logits produced prior to softmax normalization. The initial probability distribution is computed as:

$$p_i = \frac{\exp(z_i)}{\sum_{j=1}^{K} \exp(z_j)} \tag{3}$$

The predicted label is obtained as:

$$\hat{y} = \arg\max_i \; p_i \tag{4}$$

Although softmax probabilities provide an initial estimate of prediction confidence, they are often poorly calibrated and may not accurately reflect true likelihoods. Consequently, these probabilities are treated as preliminary estimates and are subsequently refined through calibration and uncertainty modeling techniques. The calibrated probablity distribution obtained from the logits is then utilized for uncertainty quantification and served as a key input for downstream reliability-aware decision-making within the proposed framework.

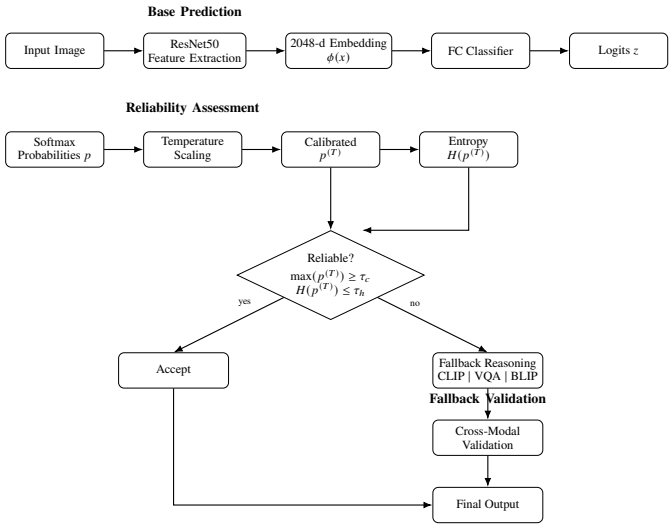

Fig. 2: Overview of the proposed reliability-aware vision-language inference framework. The pipeline consists of base prediction, reliability assessment through confidence calibration and uncertainty estimation, and fallback validation for robust final decision-making.

E. Temperature Scaling for Calibration

Address the prevalence of overconfident predictions ,we utilize temperature scaling, a post-hoc calibration methodology designed to refine the distribution of softmax outputs.The result calibrated probabilities are formulated as follows:

$$p_i^{(T)} = \frac{\exp\left(\frac{z_i}{T}\right)}{\sum_{j=1}^{K} \exp\left(\frac{z_j}{T}\right)} \tag{5}$$

In this expression, $T > 0$ denotes a learnable scalar temperature parameter,which is typically optimized using a held-out validation set to ensure effective calibration. When $T > 1$, the probability distribution is softened, thereby mitigating overconfident

output scores.Conversely, values if $T < 1$ result in a sharpening of the distribution across categories. Notably,temperature scaling maintains the original predictive ranking of the model while significantly enhancing the alignment between the predicted confidence levels and the observed empirical accuracy.

F. Uncertainly Quantification

In addition to confidence estimation,we incorporate entropy-based uncertainty modeling to assess prediction variability. The predictive entropy,derived from the calibrated probability distribution p,is formulated as follows:

$$H(p) = -\sum_{i=1}^{K} p_i \log p_i \tag{6}$$

In this context,entropy serves as a quantitative measure of precipitin uncertainty:Low entropy indicates a confident and peaked distribution.High entropy signifies uncertain and diffuse predictions. This metric facilitates the identification of unreliable outputs and provides essential data for downstream decision-making, particularly within safety-critical environments.

G. Calibration Evaluation Metrics

To rigorously assess the efficacy of the calibration process, we employ two conventional metrics that quantify the alignment between predictive confidence and empirical correctness. **Excepted Calibration Error(ECE):** The ECE serves as a quantitative measure that characterize the discrepancy between predicted confidence levels and observed accuracy by partitioning the predicting space into discrete bins as follows:

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)| \tag{7}$$

In this formulation,B_m represents the subset of samples assigned to the m-th bin,while N denotes the aggregate number of samples within the evaluation set. **Reliability Diagrams:** Reliability Diagrams provide a visual representation of calibration performance by plotting predicted confidence against empirical accuracy.Ideally, the data points align with the diagonal identify line;conversely, significant deviations from this trajectory reveal systemic overconfidence or under confidence.

Collectively,these evaluative frameworks offer both qualitative insights into robustness and reliability of model's predictive uncertainty.

H. Confidence-Aware Fallback Mechanism

A central contribution of the work is the introduction of confidence-aware fallback mechanism designed to enhance robustness under low-contribution conditions.When the calibrated confidence falls below a predefined threshold(e.g.0.60),the system activates an auxiliary reasoning pipeline: **CLIP-based zero-shot classification** is used to re-evaluate predictions.**Vision-language querying(VQA-style)** is employed to validate outputs through semantic reasoning.**Image-text similarity scoring** is used ensure cross modal consistency. Additionally,**BLIP-based captioning** is incorporated to generate descriptive textual representations, which are evaluated using both token-level probabilities and semantic similarity scores. This multi-stage fallback process enables the system to **cross-validate predictions across modalities,** thereby reducing the likelihood of incorrect high-confidence outputs.

I. Multi-modal Decision Integration

The final prediction is obtained by integrating outputs from multiple modalities along with their calibrated confidence and uncertainly measures.The system produces:Predicted label, Calibrated confidence score,Entropy-based uncertainty, Fallback validated outcome This holistic integration ensures that predictions are not only accurate but also interpretable and trustworthy,which is essential for deployment in decision-critical applications.

J. Implementation Considerations

The Framework is implemented using PyTorch-based models with modular pipelines for each modality.Efficient preprocessing strategies, including uniform sampling and batch-wise inference ,are employed to balance computational sampling and batch-wise inference,are employed to balance computational coast and performance. Despite the added complexity of calibration and fallback mechanisms, the system maintains practical inference efficiency suitable for real-world applications.

IV. Experiments and Result

A. Experimental Setup

To evaluate the effectiveness of the proposed confidence-aware multimodal framework, we conduct experiment across image and video modalities. **Image Modality:**CIFAR-100 dataset processed using a ResNet backbone.**Video Modality:** Action recognition dataset processed using R3D-18 3D CNN. **Fallback Models:** CLIP(zero-shot classification) and BLIP (image captioning).**Calibration Method:** Temperature scaling applied to logits.**Evaluation Metrics:** Expected Calibration Error(ECE),Reliability Diagrams, Prediction confidence consistency,Entropy-based uncertainty.

The evaluation focuses on both predictive performance and confidence reliability, ensuring that model outputs are not only accurate but also trustworthy.

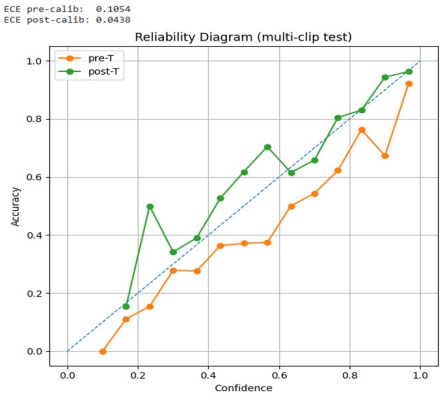

Fig. 3: Calibration improvement via temperature scaling

B. Effect of Temperature Scaling on Calibration

We first analyze the impact of temperature scaling on model calibration.

Before calibration, the models exhibits **systematic overconfidence**,where predicted probabilities significantly exceed empirical accuracy.This is evident in reliability diagrams, where data points lie consistently below the diagonal, indicating inflated confidence estimates.After applying temperature scaling , the confidence distribution becomes more aligned with true correctness.Specifically:Confidence values becomes less extreme and more representative of actual performance.The reliability curve shifts closer to the ideal diagonal.Overconfidence is significantly reduced Quantitatively, the Expected Calibration Error(ECE) improves from: $ECE_{\text{before}} = 0.1054 \rightarrow ECE_{\text{after}} = 0.0438$.This substantial reduction confirms that temperature scaling effectively enhances the reliability of model predictions.Show orange(before) vs green(after)

C. Confidence-Accuracy Alignment Analysis

To further validate calibration quality, we analyze the relationship between predicted confidence and actual accuracy. An uncalibrated model typically exhibits overconfident behavior , predicting high confidence values(e.g,95%) while the actual correctness is significantly lower (approx 80%),thereby indicating misleading certainty. In contrast, a calibrated model produces confidence estimates that closely align with observed accuracy. For instance , a prediction with 85% confidence corresponds to an actual correctness of approx 85%, demonstrating improved reliability and trustworthiness of the model's probabilistic outputs.

The alignment demonstrates that the proposed calibration pipeline transforms confidence scores into meaningful probabilistic estimates, which are critical for decision-making systems.

D. Entropy-Based Uncertainly Evaluation

We evaluate the role of entropy as a measure of predictive uncertainty.Low entropy Predictions correspond to sharply peaked distributions and high confidence.High entropy Predictions indicates ambiguous or uncertain outputs.

Entropy serves as an effective indicator for detecting unreliable predictions ,particularly in cases where confidence alone may be insufficient.Furthermore, entropy is used as a trigger signal for activating the fallback mechanism,enabling adaptive system behavior under uncertainty.

E. Multi modal Confidence Consistency

We assess whether the proposed framework maintains consistent confidence estimates across modalities. In video classification tasks, the system produces:Predicted action label,Calibrated confidence score.Top-k probability distribution,entropy-based uncertainty.

A representative example demonstrates: Predicted class: Table Tennis shot and Confidence score:0.9152 and Entropy:low

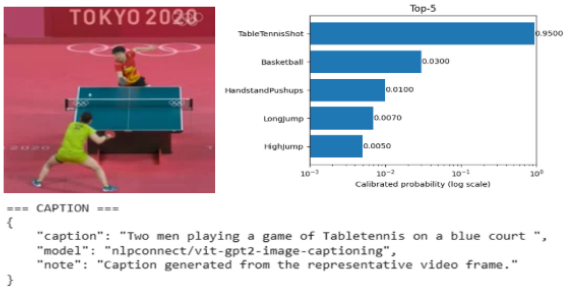

Fig. 4: This Show structured output and highlight consistency.

(indicating high certainty) To validate this prediction,a VQA-based query is issued ,yielding: Response:"yes" and Confidence:consistent with primary prediction. This agreement across classification and reasoning modules highlight the cross-modal consistency of the proposed system.

F. Effectiveness of confidence-Aware Fallback Mechanism

We evaluate the robustness of the fallback mechanism under low-confidence conditions. When the primary model produces confidence below a predefined threshold(e.g.,0.60):CLIP performs zero-shots classification.VQA queries validate semantic correctness.BLIP generates captions for additional verification.This multi-stage validation provides:Redundancy in decision-making ,cross-modal verification.Reduced likelihood of incorrect predictions.

Empirical observations indicate that the fallback mechanism: corrects uncertain predictions.improves reliability in ambiguous cases,Enhances interoperability through textual explanations.

G. Image Captioning Analysis

We further evaluate confidence estimation in image captioning using BLIP. The final caption confidence is computed as a combination of, Token-level log probabilities,Image-text similarity scores(via CLIP). As shown in Fig.5, the generated caption is "A dolphin jumping out of the water" and we got the confidence score 0.99. This demonstrates that the system is capable of assigning **quantifiable confidence to generative outputs** , extending beyond classification tasks.

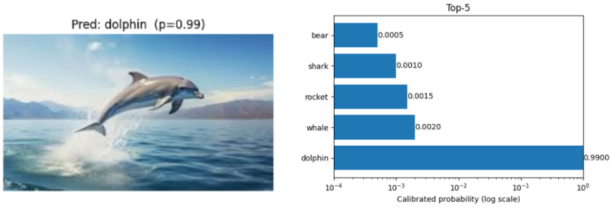

Fig. 5: Image Caption + Confidence Visualization

H. Top-K probability Distribution Analysis

To better understand prediction behavior, we analyzer the Top-5 probability distribution after calibration. The distribution becomes smother and less peaked.secondary classes retain

meaningful probabilities.Improves interpretability of model outputs. This reflects a more realistic representation of model uncertainty and reduces the risk of overconfident misclassifications.

V. Limitation

The Proposed framework has several limitations. It is highly sensitive to LLM prompt design small variations in phrasing can lead to inconsistent representation and affect downstream performance. The fallback mechanism uses a fixed confidence threshold, which is manually chosen and does not adapt to input complexity or modality specific uncertainty. Temperature scaling relies on a global parameter, which is insufficient for multimodal settings with heterogeneous uncertainty, reducing calibration robustness under distributions shifts. Dependence on multiple external models(e.g. CLIP, BLIP, VQA) increases latency and introduces error propagation, as these components are not jointly optimized.

VI. Discussion and Conclusion

The experimental results demonstrate that the proposed modality-agnostic framework effectively addresses key limitations of existing multimodal systems, particularly with respect to representation quality and confidence reliability. The integration of temperature scaling significantly improves the alignment between predicted confidence and empirical accuracy, resulting in more reliable and well-calibrated predictions. Entropy-based uncertainty modeling, combined with confidence estimation, enables the identification of ambiguous predictions and provides a more comprehensive characterization of uncertainty beyond single-point confidence measures. Furthermore, the incorporation of a confidence-aware fallback mechanism enhances system robustness by validating low-confidence predictions through cross-modal reasoning, thereby reducing erroneous outputs. The framework ensures multimodal consistency by learning semantically aligned representations across text, image, and video modalities, enabling coherent reasoning. In contrast to conventional image-centric approaches, the proposed method produces calibrated and semantically grounded predictions, making it suitable for decision-critical applications. Overall, the approach achieves its objective of integrating semantic alignment, confidence calibration, uncertainty quantification, and robust validation into a unified system, leading to improved performance and trustworthiness. Future work may focus on improving interpretability through feature-level alignment analysis and extending the framework to support multi-class prediction. Additional directions include adaptive calibration strategies and scalable methods for real-time multimodal inference.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PmLR, 2021.

[2] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "Imagebind: One embedding space to bind them all," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15180–15190, 2023.

[3] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "Audioclip: Extending clip to image, text and audio," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 976–980, IEEE, 2022.

[4] H. Xu, G. Ghosh, P.-Y. Huang, D. Okhonko, A. Aghajanyan, F. Metze, L. Zettlemoyer, and C. Feichtenhofer, "Videoclip: Contrastive pre-training for zero-shot video-text understanding," in *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 6787–6800, 2021.

[5] R. Zhang, Z. Guo, W. Zhang, K. Li, X. Miao, B. Cui, Y. Qiao, P. Gao, and H. Li, "Pointclip: Point cloud understanding by clip," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8552–8562, 2022.

[6] Y. Lyu, X. Zheng, J. Zhou, and L. Wang, "Unibind: Llm-augmented unified and balanced representation space to bind them all," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26752–26762, 2024.

[7] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International conference on machine learning*, pp. 1321–1330, PMLR, 2017.

[8] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," *Advances in neural information processing systems*, vol. 32, 2019.

[9] F. Liu, K. Lin, L. Li, J. Wang, Y. Yacoob, and L. Wang, "Mitigating hallucination in large multi-modal models via robust instruction tuning," *arXiv preprint arXiv:2306.14565*, 2023.

[10] Y. Geifman and R. El-Yaniv, "Selectivenet: A deep neural network with an integrated reject option," in *International conference on machine learning*, pp. 2151–2159, PMLR, 2019.

[11] Y.-J. Sun, S. Dey, D. Hakkani-Tür, and G. Tur, "Confidence estimation for llm-based dialogue state tracking," in *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 1083–1090, IEEE, 2024.

[12] S. E. Spatharioti, D. Rothschild, D. G. Goldstein, and J. M. Hofman, "Effects of llm-based search on decision making: Speed," *Accuracy, and Overreliance*, 2025.