Unified Multimodal Uncertain Inference

Anonymous ACL submission

Abstract

We introduce Unified Multimodal Uncertain Inference (UMUI), a multimodal inference task spanning text, audio, and video, where models must produce calibrated probability estimates of hypotheses conditioned on a premise in any modality or combination. While uncertain inference has been explored in text, extension to other modalities has been limited to single-modality binary entailment judgments, leaving no framework for fine-grained probabilistic reasoning in or across other modalities. To address this, we curate a human-annotated evaluation set with scalar probability judgments across audio, visual, and audiovisual settings, and additionally evaluate on existing text and audio benchmarks. We introduce CLUE, which combines self-consistent teacher calibration and distribution-based confidence probing to produce well-calibrated predictions. We demonstrate that our 3B-parameter model achieves better or similar performance than baselines up to 32B parameters across all modalities.[1]

1 Introduction

Human reasoning is fundamentally probabilistic, operating over likelihoods, degrees of belief, and confidence estimates rather than categorical or binary judgments (Oaksford and Chater, 2001; Chater et al., 2006; Tenenbaum et al., 2011). When viewing social media footage in the aftermath of an earthquake, we integrate the extent of visible structural collapse, the sounds of sirens and car alarms, and bystanders' reactions to form a graded estimate of severity – not a binary classification, but a calibrated belief synthesized across vision and audio. More broadly, this probabilistic reasoning operates fluidly across modalities, with beliefs continuously updated as we read, watch, and listen. As multimodal AI systems are increasingly deployed in high-stakes domains, their inability to express calibrated uncertainty over multimodal evidence limits their trustworthiness and utility.

Uncertain Natural Language Inference (UNLI) addresses this gap in text, replacing categorical entailment labels with scalar probability estimates that better reflect human judgment (Chen et al., 2020). However, uncertainty-aware inference remains almost exclusively textual (Geng et al., 2024; Xiong et al., 2024; Wang et al., 2025). Extensions to other modalities remain limited: in vision, calibrated uncertainty has been explored for event classification (Sanders et al., 2022), while inference tasks in audio and video have been formulated only as binary entailment judgments (Drossos et al., 2020; Liu et al., 2020; Deshmukh et al., 2024; Martin et al., 2025b). As a result, no existing framework supports general fine-grained probabilistic inference in or across non-textual modalities.

We address these limitations by introducing Unified Multimodal Uncertain Inference (UMUI), a task for calibrated probabilistic inference spanning text, audio, video, and their combinations. To support this task, we collect human scalar probability judgments for audio-visual premise-hypothesis pairs. We find that existing multimodal language models are poorly calibrated for this task, motivating CLUE (**C**alibrated **L**atent **U**ncertainty **E**stimation), a training approach that combines self-consistent teacher calibration, distribution-based confidence probing, and modality-specific batching to produce well-calibrated predictions from a single multimodal model.

Our contributions are: (1) We formalize UMUI, the task of producing calibrated probability estimates of hypotheses conditioned on a premise in any modality or combination; (2) We collect human scalar probability judgments for audio-visual premise-hypothesis pairs, extending existing binary annotations to fine-grained uncertainty labels; and (3) We introduce CLUE and demonstrate that our 3B-parameter model achieves better performance

[1] We will release the data and code upon publication.

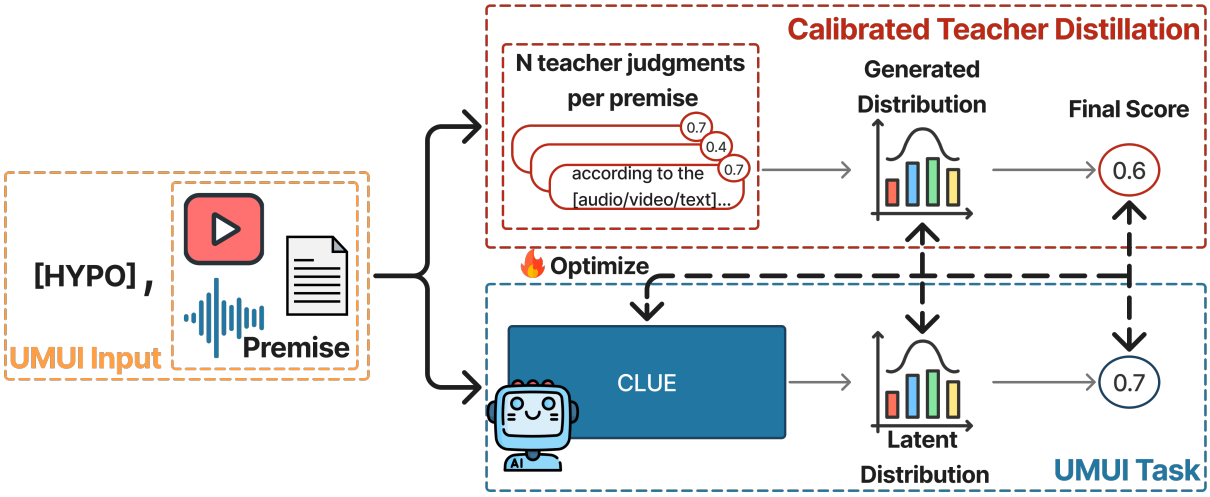

Figure 1: Overview of UMUI and CLUE. UMUI takes a hypothesis and multimodal premise (audio, video, and/or text) and produces a calibrated probability estimate of the hypothesis. **Top right:** To generate training labels, we produce N independent teacher judgments per premise-hypothesis pair and aggregate them into calibrated distillation scores. **Bottom right:** CLUE takes a hypothesis-premise pair in any modality and predicts a latent distribution over confidence tokens, which is aggregated into a scalar probability estimate.

than baselines up to 32B parameters across text, audio, visual, and audiovisual settings.

2 UMUI

Task Definition The UMUI task takes as input a *premise*, multimodal information such as a sentence, audio clip, or video, and a *hypothesis*, a declarative statement that may or may not be supported by the premise. The model produces a scalar *prediction* in [0, 1] representing the probability that the hypothesis is true given the premise (Figure 1).

Data Collection Since Chen et al. (2020) provide scalar probability annotations for text, we focus on annotating audio, visual, and audiovisual premises. We build on WIKIVIDEO (Martin et al., 2025a), which provides binary support judgments for claims (serving as hypotheses) against video content (serving as premises) in the MULTIVENT2.0 test collection (Kriz et al., 2025). We select 10 topics, following Martin et al. (2025b), for our test set and for each video-claim pair, we elicit scalar probability judgments from expert annotators. Annotators also indicate which modalities (audio, video, or both) informed their judgment. We provide details on the annotation in Appendix B.

3 CLUE

Our method consists of three components: (1) *self-consistent teacher calibration*, where we aggregate multiple reasoning traces from a teacher model to produce calibrated training labels; (2) *latent distribution confidence modeling*, where we probe a model's internal latent distribution; and (3) *modality-specific batching*, which groups training samples by modality to improve gradient stability and reduce padding overhead.

3.1 Self-Consistent Teacher Calibration

To address the lack of fine-grained probability judgments in the CLOTHO (Drossos et al., 2020) and WIKIVIDEO datasets, we employ a teacher model to generate labels. However, as shown in Table 1, teacher models are poorly calibrated in zero-shot. Following prior work in self-consistency (Wang et al., 2023b; Xiong et al., 2024), we calibrate the teacher without training by having it produce multiple reasoning traces. For each premise-hypothesis pair, the teacher produces five independent responses, which are aggregated by either mean (average model confidence) or max (majority model confidence). As shown in Table 1, both methods increase accuracy (binary) and reduce MSE, with mean showing the greatest improvement. We note that the results in Table 1 are evaluated on a smaller test set redundantly annotated by all annotators; further details on teacher selection are in Appendix C.

3.2 Latent Distribution Confidence

To produce calibrated probability estimates, we design a model that predicts a latent distribution over confidence levels rather than generating a scalar value as discrete tokens. Inspired by Wang et al. (2025), we discretize the interval $[0, 1]$ into

Model	Votes	Acc.	MSE
Q3-VL	1	0.6571	3.673e-2
	5-max	0.8571	8.748e-4
	5-mean	0.8571	1.560e-5
Q3-O	1	0.7333	1.004e-2
	5-max	0.7619	5.863e-5
	5-mean	0.7619	2.404e-5

Table 1: Accuracy and mean squared error for self-consistency. Q: Qwen, VL-Vision Language, O-Omni, both thinking models. Votes: 1 (single inference), 5-max (majority vote), 5-mean (average probability).

N bins $\{b_0, b_1, \ldots, b_{N-1}\}$ and associate each bin with a dedicated confidence token <CONF_i>. During training, the model learns to predict a distribution over these tokens, supervised by a target Gaussian distribution $Q \sim \mathcal{N}(y, \sigma^2)$, where y is the ground-truth probability label and σ is set to a small value to approximate a sharply peaked distribution. We train this model by minimizing the Kullback–Leibler divergence (Kullback and Leibler, 1951) between the target distribution Q and the predicted distribution P: $\mathcal{L}_d = D_{\mathrm{KL}}(Q \| P)$. In the experiment, we set N to 100 and σ to 0.05.

At inference time, a smooth probability estimate is reconstructed via weighted aggregation:

$$\hat{p} = \sum_i f(b_i)\, p(\texttt{<CONF_i>}), \qquad (1)$$

where $f(\cdot)$ maps each bin b_i to its representative probability value. With this, we are able to capture the model's internal confidence distribution enabling smoother estimates than discrete token outputs or logit sampling.

We compare this approach against a **token-based baseline**, where the model directly generates a probability value as natural language tokens. Since the output is produced autoregressively, the resulting probabilities are inherently discrete (e.g., 0.5, 0.8), limiting the model's ability to express fine-grained confidence. The token-based model is trained with a standard autoregressive objective.

3.3 Training

To optimize training efficiency and gradient stability, we implement a modality-specific batching strategy, grouping samples of the same modality into uniform batches rather than constructing heterogeneous batches of text, audio, and video. This approach aligns with the balance-aware strategy in Li et al. (2025). Mixing modalities within a batch leads to unbalanced gradient magnitudes, which can cause numerical instability or gradient interference during optimization (Chen et al., 2018). Additionally, the significant length disparity across modalities (e.g., $L_{\text{text}} \ll L_{\text{video}}$) results in excessive padding that dilutes the gradient signal for shorter modalities (Dao et al., 2022). Modality-specific batching addresses both issues.

Model	P	Clot	WV-V	WV-A	WV-AV
AF	7B	48.9	N/A	18.7	N/A
Q2-A	7B	48.7	N/A	34.7	N/A
Q2.5-VL	32B	N/A	22.1	N/A	N/A
Q3-VL	32B	N/A	**81.1**	N/A	N/A
Q2.5-O	7B	50.1	50.5	66.6	45.2
CLUE-T	3B	**97.5**	56.3	**71.5**	54.3
CLUE-D	3B	95.8	74.6	70.2	**70.1**

Table 2: Accuracy of binary judgments. P: Parameters. AF: Audio Flamingo. Q: Qwen, A-Audio VL-Vision Language, O-Omni. Clot: Clotho, WV: WIKIVIDEO, V-Vision only, A-Audio only, AV-Audio-Visual.

4 Experiments

Training We initialize our models from Qwen2.5-Omni-3B (Xu et al., 2025) and report specific training parameters in Appendix D.

Evaluation Setup We evaluate on three datasets: CLOTHO (binary audio, Drossos et al., 2020; Deshmukh et al., 2024), UNLI (scalar text, Chen et al., 2020), and WIKIVIDEO claims for binary and scalar judgments in audio, visual, and audiovisual settings. Baselines span four classes: (1) audio-only: Audio Flamingo 3 (AF, Goel et al., 2025) and Qwen2-Audio (Q2-A, Chu et al., 2024); (2) language-only: Qwen3 (Q3, Yang et al., 2025); (3) vision-language: Qwen2.5-VL (Q2.5-VL, Bai et al., 2025b) and Qwen3-VL (Q3-VL, Bai et al., 2025a); and (4) omnimodal: Qwen2.5-Omni (Q2.5-O, Xu et al., 2025). For binary evaluation, we report accuracy; for scalar evaluation, we report mean squared error (MSE). Baseline models generate yes/no judgments for binary settings and scalar judgments via an optimized prompt (Appendix G). For our scalar models, we apply a naive 0.5 threshold for binary decisions.

Binary Judgments In Table 2, we report the accuracy of binary judgments on CLOTHO and

Model	P	UNLI	WV-V	WV-A	WV-AV
AF	7B	27.2	N/A	49.9	N/A
Q2-A	7B	77.8	N/A	62.6	N/A
Q3	32B	10.6	N/A	N/A	N/A
TMTO	14B	7.5	N/A	N/A	N/A
Q2.5-VL	32B	12.6	13.3	N/A	N/A
Q3-VL	32B	7.5	10.5	N/A	N/A
Q2.5-O	7B	11.4	14.1	14.3	15.2
CLUE-T	3B	8.1	10.1	**3.4**	9.8
CLUE-D	3B	**5.7**	**7.8**	8.4	**7.9**

Table 3: Mean squared error (x100) for UMUI judgments. P: Parameters. AF: Audio Flamingo. Q: Qwen, A-Audio VL-Vision Language, O-Omni. WV: WIKIVIDEO, V-Vision only, A-Audio only, AV-Audio-Visual. TMTO: model from Wang et al. (2025).

WIKIVIDEO (binary). On CLOTHO, CLUE substantially outperforms audio baselines, from near chance to near perfect performance. On WIKIVIDEO, results vary by modality, however, we see the same trend that CLUE outperforms all methods in any modality or combination.

Scalar Judgments In Table 3, we report the MSE for scalar judgments on UNLI and WIKIVIDEO (scalar). On UNLI, our distribution-based model achieves the lowest MSE, outperforming text-only models up to 32B parameters. Across all WIKIVIDEO settings, both of our 3B models outperform the omnimodal and modality specific baselines, with the largest gains in the vision-only and audiovisual settings. These results demonstrate that CLUE is well-calibrated across modalities.

Token vs Distribution Comparing the two model variants, token-based (T) and distribution-based (D), we see that CLUE-D generally achieves lower and more stable performance, while CLUE-T seems to perform better at a single modality (audio). We attribute the overall advantage of the distribution-based model to its ability to express smooth probability estimates via a latent distribution over confidence tokens, whereas the token-based model is constrained to discrete outputs.

Training Mixture Table 4 compares training ratios of positive and negative examples. The 1:1 ratio achieves the lowest MSE, while increasing negative samples degrades both metrics. We select the 1:1 ratio for its balance between calibration and categorical performance. Visualizations of predicted distributions are in Appendix E and show a negative shift in expectation for unbalanced data.

Ratio	MSE	Acc.	F1
1:0.5	0.1365	0.794	0.877
1:1	0.0797	0.701	0.811
1:2	0.0931	0.642	0.744
1:3	0.0820	0.469	0.553
1:6.5 (full)	0.1030	0.438	0.472
1:1 (token)	0.0978	0.543	0.665

Table 4: Comparison of different (pos:neg) training ratios on performance for WIKIVIDEO-AV (scalar).

Calibrated Comparison In Table 3, we also compare against a modality-specific calibrated method (TMTO, Wang et al., 2025), which uses a similar latent distribution approach but is trained exclusively on text. Despite being trained on a mixture of modalities at 3B parameters, CLUE-D outperforms TMTO (14B) on UNLI, suggesting that signal from the other modalities helps improve performance in text. This result supports our hypothesis that multimodal uncertain inference can benefit single modality applications.

5 Related Work

Well-calibrated confidence signals enable efficient resource allocation, from deferring to human experts (Jurayj et al., 2025) to early-stopping policies that reduce inference costs (Wang et al., 2026). At the token level, logit-based calibration has been applied to video reranking (Skow et al., 2026), while self-consistency (Wang et al., 2023b) and LoRA ensembles (Wang et al., 2023a) offer lightweight alternatives to deep ensembles (Lakshminarayanan et al., 2017) for mitigating overconfidence (Xiong et al., 2024). Multimodal uncertainty quantification remains nascent, largely limited to images and relying on costly inference-time methods (Guo et al., 2017; Kadavath et al., 2022; Chen et al., 2025). We provide additional references in Appendix A.

6 Conclusion

We introduce UMUI, the task of calibrated probabilistic inference across text, audio, video, and their combinations. We propose CLUE, combining self-consistent teacher calibration, latent distribution confidence modeling, and modality-specific batching, and show that our 3B-parameter model achieves lower calibration error than baselines up to 32B parameters across all tested modalities.

Limitations

Our evaluation set, while carefully annotated, is limited in scale: we collect scalar probability judgments on 10 topics from 4 annotators. Additionally, our training data relies entirely on teacher-generated labels rather than human-annotated scalar probabilities, as collecting fine-grained probability judgments at training scale is prohibitively expensive. While we show that self-consistent teacher calibration produces labels well-aligned with human judgment, the quality ceiling of our trained models is ultimately bounded by the fidelity of these synthetic labels.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Christopher M Bishop and Nasser M Nasrabadi. 2006. *Pattern recognition and machine learning*, volume 4. Springer.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Nick Chater, Joshua B. Tenenbaum, and Alan Yuille. 2006. Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, 10(7):287–291. Special issue: Probabilistic models of cognition.

Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. 2020. Uncertain natural language inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8772–8779, Online. Association for Computational Linguistics.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. 2018. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pages 794–803. PMLR.

Zijun Chen, Wenbo Hu, Guande He, Zhijie Deng, ZHeng ZHang, and Richang Hong. 2025. Unveiling uncertainty: A deep dive into calibration and performance of multimodal large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3095–3109, Abu Dhabi, UAE. Association for Computational Linguistics.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *Preprint*, arXiv:2407.10759.

Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. 2025. Beyond binary rewards: Training lms to reason about their uncertainty. *arXiv preprint arXiv:2507.16806*.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359.

Soham Deshmukh, Shuo Han, Hazim Bukhari, Benjamin Elizalde, Hannes Gamper, Rita Singh, and Bhiksha Raj. 2024. Audio entailment: Assessing deductive reasoning for audio understanding. *Preprint*, arXiv:2407.18062.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. *Preprint*, arXiv:2311.08298.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *Preprint*, arXiv:2507.08128.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

William Jurayj, Jeffrey Cheng, and Benjamin Van Durme. 2025. Is that your final answer? test-time scaling improves selective question answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 636–644, Vienna, Austria. Association for Computational Linguistics.

William Jurayj, William Rudman, and Carsten Eickhoff. 2022. Garden path traversal in GPT-2. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 305–313, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Reno Kriz, Kate Sanders, David Etter, Kenton Murray, Cameron Carpenter, Kelly Van Ochten, Hannah Recknor, Jimena Guallar-Blasco, Alexander Martin, Ronald Colaianni, Nolan King, Eugene Yang, and Benjamin Van Durme. 2025. Multivent 2.0: A massive multilingual benchmark for event-centric video retrieval. *Preprint*, arXiv:2410.11619.

S. Kullback and R. A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Zhiqi Li, Guo Chen, Shilong Liu, Shihao Wang, Vibashan VS, Yishen Ji, Shiyi Lan, Hao Zhang, Yilin Zhao, Subhashree Radhakrishnan, and 1 others. 2025. Eagle 2: Building post-training data strategies from scratch for frontier vision-language models. *arXiv preprint arXiv:2501.14818*.

Jianhua Lin. 2002. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Jingzhou Liu, Wenhu Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang, and Jingjing Liu. 2020. Violin: A large-scale dataset for video-and-language inference. *Preprint*, arXiv:2003.11618.

Alexander Martin, Reno Kriz, William Gantt Walden, Kate Sanders, Hannah Recknor, Eugene Yang, Francis Ferraro, and Benjamin Van Durme. 2025a. Wikivideo: Article generation from multiple videos. *Preprint*, arXiv:2504.00939.

Alexander Martin, William Walden, Reno Kriz, Dengjia Zhang, Kate Sanders, Eugene Yang, Chihsheng Jin, and Benjamin Van Durme. 2025b. Seeing through the mirage: Evaluating multimodal retrieval augmented generation. *Preprint*, arXiv:2510.24870.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

Mike Oaksford and Nick Chater. 2001. The probabilistic approach to human reasoning. *Trends in cognitive sciences*, 5 8:349–357.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Karl Pearson. 1895. Vii. note on regression and inheritance in the case of two parents. *proceedings of the royal society of London*, 58(347-352):240–242.

Kate Sanders, Reno Kriz, Anqi Liu, and Benjamin Van Durme. 2022. Ambiguous images with human judgments for robust visual event classification. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Tyler Skow, Alexander Martin, Benjamin Van Durme, Rama Chellappa, and Reno Kriz. 2026. Rankvideo: Reasoning reranking for text-to-video retrieval. *Preprint*, arXiv:2602.02444.

Charles Spearman. 1961. The proof and measurement of association between two things.

Elias Stengel-Eskin, Kyle Rawlins, and Benjamin Van Durme. 2024. Zero and few-shot semantic parsing with ambiguous inputs. In *The Twelfth International Conference on Learning Representations*.

Joshua B. Tenenbaum, Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. 2011. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285.

Uddeshya Upadhyay, Shyamgopal Karthik, Massimiliano Mancini, and Zeynep Akata. 2023. Probvlm: Probabilistic adapter for frozen vison-language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1899–1910.

Liaoyaqi Wang, Zhengping Jiang, Anqi Liu, and Benjamin Van Durme. 2025. Always tell me the odds: Fine-grained conditional probability estimation. *arXiv preprint arXiv:2505.01595*.

Xi Wang, Laurence Aitchison, and Maja Rudolph. 2023a. Lora ensembles for large language model fine-tuning. *arXiv preprint arXiv:2310.00035*.

Xi Wang, Anushri Suresh, Alvin Zhang, Rishi More, William Jurayj, Benjamin Van Durme, Mehrdad Farajtabar, Daniel Khashabi, and Eric Nalisnick. 2026. Conformal thinking: Risk control for reasoning on a compute budget. *arXiv preprint arXiv:2602.03814*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

A Extended Related Work

A.1 Uncertainty Quantification and Calibration of Language Models

Reliable uncertainty estimation is essential for the trustworthy deployment of multimodal language models. Recent work has demonstrated that well-aligned confidence signals enable more efficient allocation of resources, either through deferring to human experts in high-risk settings (Jurayj et al., 2025) or to define early-stopping policies to minimize inference costs (Wang et al., 2026). These approaches differ in where and how uncertainty is measured. At the token level, Skow et al. (2026) show that logit-based calibration can be used to rerank videos by relevance, while ensemble methods like self-consistency (Wang et al., 2023b) or LoRA ensembles (Wang et al., 2023a) offer lightweight adaptations of deep ensembles (Lakshminarayanan et al., 2017) to help mitigate language models' tendency towards overconfidence (Xiong et al., 2024), and reasoning can help surface these latent signals into verbalized uncertainty scores (Damani et al., 2025). Research into multimodal uncertainty quantification remains nascent, focusing on images and relying on costly inference-time policies (Chen et al., 2025), implementing popular methods in neural network (Guo et al., 2017) and language model (Kadavath et al., 2022). In contrast, we show how multimodal models can directly learn to convey their uncertainty when processing dynamic video inputs.

A.2 Fine Grained Probability Estimation

Conditional probability estimation (Chen et al., 2020) extends traditional natural language inference (Bowman et al., 2015) to measure language models' ability to evaluate the likelihood of a hypothesis given a premise on a continuous scale. This setting extends conventional binary judgments toward metrics of credence that more closely align with human uncertainty (Pavlick and Kwiatkowski, 2019; Nie et al., 2020). Although language models naturally struggle to process ambiguities evident to humans (Jurayj et al., 2022; Stengel-Eskin et al., 2024), targeted interventions such as synthetic data supervision can help mitigate this pathology (Wang et al., 2025). Although entailment (Xie et al., 2019) and uncertainty quantification (Upadhyay et al., 2023) have been independently studied for multimodal problems, this line of work has been conducted almost exclusively in the text and image domain. Our work extends the uncertain inference to video-language models, showing how models learn to integrate visual, auditory, and linguistic cues into well-calibrated probabilistic judgments.

B Annotation Protocol

B.1 Data Selection

We select the same events from WIKIVIDEO as Martin et al. (2025b). We do this to ensure a broad coverage of event types, dates (ensuring some are out of model parametric knowledge), and to allow for our method and others to be evaluated in applications like multimodal RAG evaluation. The full event list is:

1. Launch and commissioning of the James Webb Space Telescope
2. 2018 lower Puna eruption
3. Notre-Dame fire
4. 2022 United States Senate election in Georgia

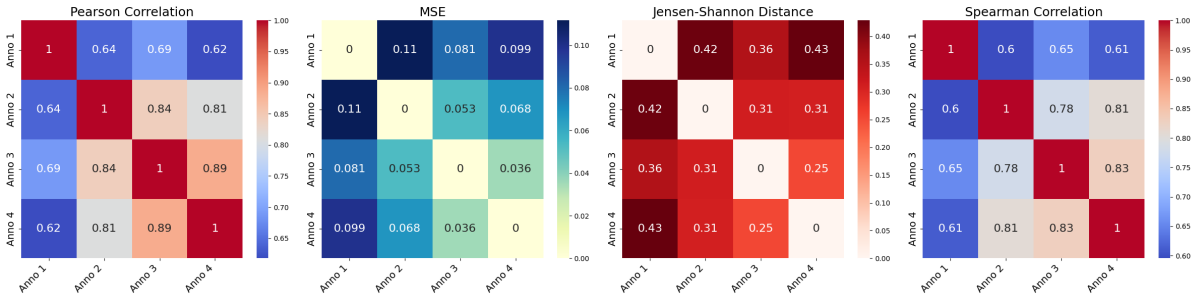

Figure 2: Annotators agreement analysis across four metrics. The heatmaps display pairwise consistency among annotators using Pearson Correlation, Mean Squared Error (MSE), Jensen-Shannon Distance, and Spearman Correlation, illustrating the reliability and alignment of human evaluations.

5. Hurricane Irma
6. 2018 Anchorage earthquake
7. 2025 Canadian federal election
8. 2025 Myanmar earthquake
9. Blue Ghost Mission 1
10. Liberation Day Tariffs

B.2 Annotation Protocol

In Figure 10, we provide the annotation instructions given to the annotators. Annotators are given a video and a set of claims and asked to estimate the probability (0–100%) that each claim is true given the video. For each claim, annotators also indicate which modalities (audio, video, or both) informed their judgment. To ensure consistency, annotators are provided with a background story for each event and instructed to treat the event as currently occurring. Annotators may use general background knowledge but are prohibited from searching for information that would directly confirm or deny a claim. The annotation interface can be seen in Figure 8 and Figure 9.

B.3 Annotation Consistency and Quality Control

The final dataset was annotated by four expert annotators. To ensure high-quality data, we first had these annotators redundantly annotate a pilot study to evaluate and ensure inter-annotator agreement. In this pilot, we found that the annotators have a Krippendorff's α (Krippendorff, 2011) of 0.7121, which indicates high agreement between the annotators. In Figure 2, we also visualize other alignment calculations between these annotators with Pearson correlation (Pearson, 1895) and Spearman's rank correlation (Spearman, 1961), Mean Squared Error (Bishop and Nasrabadi, 2006), and Jensen-Shannon distance (Lin, 2002).

B.4 Data Statistics

After verifying the consistency of the annotations, we compiled a final evaluation dataset consisting of 2,526 human-labeled video-claim pairs.

C Teacher Model Selection

We select a set of potential teacher models: Qwen3-VL-32B (Thinking and non-Thinking), Qwen3-Omni-30B (Thinking and non-Thinking). To evaluate the teacher reliability, we have these teacher models operate on the same set of redundant annotations that were used for annotator agreement. In Table 5, we compare the performance of different teacher models across three metrics. The results indicate that Thinking models significantly outperform their non-Thinking counterparts. Specifically, Qwen3-VL Thinking and Qwen3-Omni Thinking achieve substantially higher Accuracy and lower MSE. Furthermore, the Thinking versions show stronger reliability with higher Krippendorff's α values, suggesting they are more reliable as teacher models for generating high-quality annotations.

In Figure 3, the Thinking models also exhibit a sharp bimodal distribution that closely mirrors the '0-or-1' scoring pattern of human annotators. Unlike non-Thinking models, which show more uncertainty in the middle range, the Thinking models provide more decisive and human-consistent judgments, making them the superior choice for high-quality data annotation.

D Training Details

We utilize three primary datasets for training: Clotho, UNLI, and WikiVideo. Since the UNLI

Model	Votes	Acc.	MSE	Krippendorff's α
Q3-VL	1	0.6190	2.722e-2	0.5659
	5-max	0.6666	1.391e-2	0.5998
	5-mean	0.6666	1.296e-2	0.5984
Q3-VL Thinking	1	0.6571	3.673e-2	0.5893
	5-max	0.8571	8.748e-4	0.6078
	5-mean	0.8571	1.560e-5	0.6708
Q3-O	1	0.3142	2.153e-1	0.4161
	5-max	0.5333	6.801e-5	0.5790
	5-mean	0.5714	6.727e-2	0.5677
Q3-O Thinking	1	0.7333	1.004e-2	0.6136
	5-max	0.7619	5.863e-5	0.6125
	5-mean	0.7619	2.404e-5	0.6501

Table 5: Accuracy, Mean squared error and Krippendorff's α for teacher model evaluation. Q: Qwen, VL-Vision Language, O-Omni.

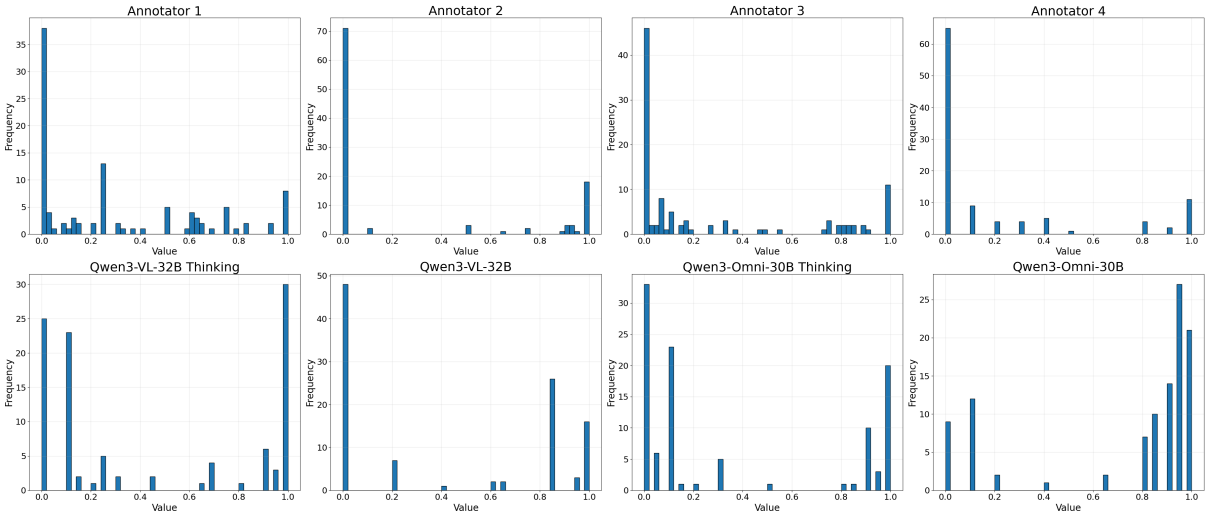

Figure 3: Probability distribution alignment between human annotators and teacher models.

Hyperparameter	Value
Number of Epochs	15
Learning Rate	1×10^{-4}
Warmup Steps	5,000
Gradient Accumulation Steps	4
LoRA Rank (r)	16
LoRA Alpha (α)	32
LoRA Dropout	0.1

Table 6: Training Hyperparameters and LoRA Configuration

dataset provides fine-grained probability distributions while Clotho and WikiVideo do not, we generated synthetic soft labels for the latter two datasets to maintain label consistency across the entire training set.

The data partitioning is as follows:

1. Clotho: Development split for training; evaluation split for testing.

2. WikiVideo: 10 events reserved for testing; the remainder for training.

3. UNLI: Standard train/validation splits for training and testing.

The model was trained on four NVIDIA A100 (80GB) GPUs using the DeepSpeed ZeRO-2 optimization stage. We utilized Low-Rank Adaptation (LoRA) for parameter-efficient fine-tuning (Hu et al., 2022). The detailed hyperparameter configurations are summarized in Table 6.

E Model Distribution

In this section, we further explore the effect of training data balance and the model's output distribution on the test set. In Figure 4, we visualize the test set predictions of each ratio. Here, we see a clear shift in the distributions as the number of negative samples increases, with the expectation shifting towards 0 with every increase in negative samples. Interestingly, we see that, with the exception of 0.0-0.1, the 1:0.5 ratio seems to best reflect the human distribution. However, without enough negative signal, the model doesn't accurately discriminate negative samples.

F Additional Results

In Table 7, we report additionally metrics (ECE and NLL) on the scalar judgments.

G Prompts

In Figure 5 (audio instances), Figure 6 (vision instances), and Figure 7 (text instances), we show the prompts designed to elicit scalar judgments from the models, enabling them to provide fine-grained probabilistic judgments.

H Acknowledgment of AI

AI assistants were used in this project for coding and to edit text for fluency and typos.

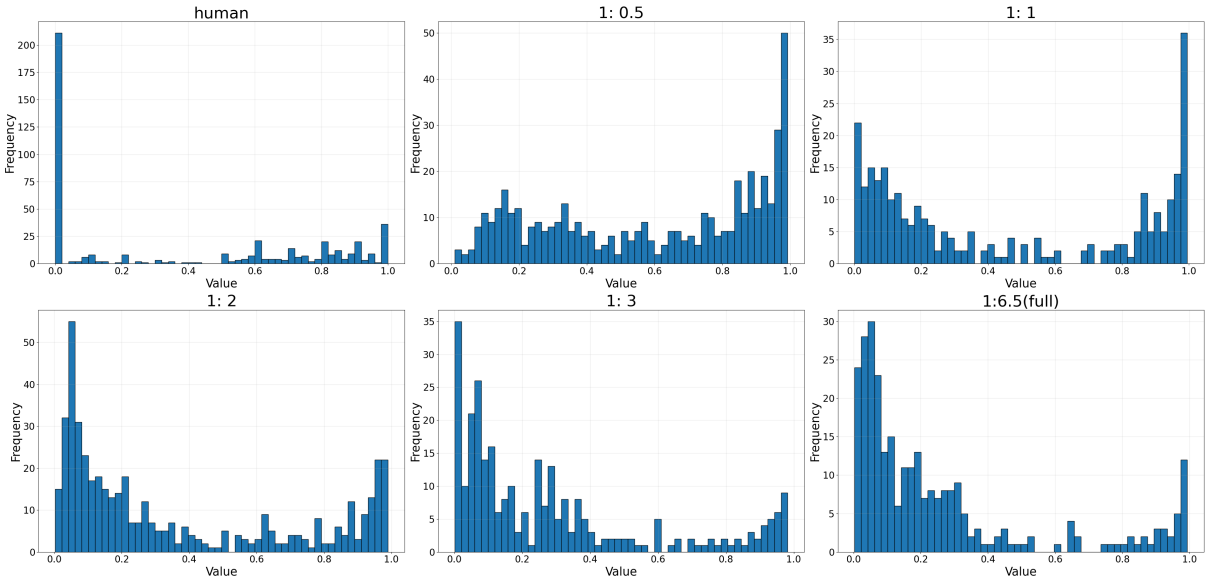

Figure 4: Probability distribution between human annotators and results from different training ratio.

		UNLI			WV-V			WV-A			WV-AV		
Model	**P**	MSE	NLL	ECE	MSE	NLL	ECE	MSE	NLL	ECE	MSE	NLL	ECE
AF	7B	2.72e-1	8.87	3.85e-1	N/A	N/A	N/A	4.99e-1	12.3	5.39e-1	N/A	N/A	N/A
Q2-A	7B	7.78e-1	8.85	3.79e-1	N/A	N/A	N/A	6.26e-1	15.3	6.68e-1	N/A	N/A	N/A
Q3	32B	1.06e-1	1.09	1.33e-1	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
TMTO	14B	7.57e-2	5.55e-1	1.47e-1	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Q2.5-VL	32B	1.26e-1	3.10	1.55e-1	1.33e-1	2.57	1.74e-1	N/A	N/A	N/A	N/A	N/A	N/A
Q3-VL	32B	7.53e-2	6.66e-1	**2.98e-2**	1.05e-1	2.71	1.15e-1	N/A	N/A	N/A	N/A	N/A	N/A
Q2.5-O	7B	1.14e-1	1.51	1.42e-1	1.42e-1	3.94	1.62e-1	1.43e-1	3.35	1.42e-1	1.52e-1	4.06	1.85e-1
CLUE-T	3B	8.10e-2	9.55e-1	9.70e-2	1.00e-1	8.01e-1	1.21e-1	**3.35e-2**	**2.30e-1**	**7.60e-2**	9.78e-2	8.72e-1	1.21e-1
CLUE-D	3B	**5.73e-2**	**5.09e-1**	5.92e-2	**7.84e-2**	**4.97e-1**	**8.47e-2**	8.40e-2	3.33e-1	1.90e-1	**7.92e-2**	**5.03e-1**	**8.42e-2**

Table 7: Calibration metrics for UMUI judgments. P: Parameters. AF: Audio Flamingo. Q: Qwen, A-Audio VL-Vision Language, O-Omni. WV: WIKIVIDEO, V-Vision only, A-Audio only, AV-Audio-Visual. TMTO: model from Wang et al. (2025). Lower is better for all metrics.

To help you make more accurate and consistent judgments, here is an expanded explanation of how to interpret and assign support percentages for audio-based evidence. These examples cover a range of real-world cases you may encounter in this annotation task.
100% - Full and unambiguous support: The audio clearly and directly contains the exact event, statement, or content described in the claim. There is no need for guessing or interpretation — the claim is fully verified by the audio.
80-100% - Almost complete support: The main content of the claim is clearly supported by the audio, though there may be minor ambiguity in speaker identity, context, or completeness (e.g., partial recording, mild noise). Overall, the claim is strongly supported.
60-80% - Strong partial support: The audio strongly suggests that the claim is true, but some key details may be missing, unclear, or ambiguous — such as incomplete phrases, background noise, or partial conversations. The evidence is strong but not definitive.
40-60% - Moderate partial support: There is some alignment between the audio and the claim, but large portions are missing, unclear, or open to interpretation. While the recording points in the same general direction as the claim, it lacks clarity or completeness for confident verification.
20-40% - Minimal weak support: There are small verbal cues or contextual hints that could relate to the claim, but they are insufficient to provide confidence in its truth.
0-20% - Very weak or speculative support: There may be the slightest indirect reference (such as a related topic or similar voice), but nothing concrete that verifies the claim.
0% - No support or contradiction: The audio does not relate to the claim at all, or it directly contradicts it.
Based on the provided audio and text, evaluate the probability that the text statement is true.
Your answer must be a decimal number between 0 and 1, and you must strictly follow the format below:
<answer>probability_value</answer>
Where probability_value is the result you calculate.
The text to evaluate is: {text}

Figure 5: Zero-Shot scaler prompt for audio examples.

To help you make more accurate and consistent judgments, here is an expanded explanation of how to interpret and assign support percentages. These examples are designed to cover a range of real-world cases you may encounter in the annotation task.
100% - /Full and unambiguous support: The video clearly shows the exact event described in the claim. There is no need for guessing or interpretation.
80-100% - Almost complete support: The main content in the claim is shown, but there may be minor ambiguity in location, identity, or completeness. The overall claims are supported by the video.
60-80% - Strong partial support: The video strongly suggests the claim is true, but some critical details may be missing, obscured, or ambiguous, limiting the ability to confirm the claim with certainty. The video gives strong but not definitive support.
40-60% - Moderate partial support: There is some alignment with the claim, but large portions are either missing, unclear, or open to interpretation. While the footage may point in the same general direction as the claim, it lacks the clarity or completeness needed for confident verification.
20-40% - Minimal weak support: There are small visual or audio cues that could hint at the claim, but they are insufficient to be confident.
0-20% - Very weak or speculative support: There may be the slightest indirect reference, such as a related object or setting, but nothing concrete happens.
0% - No support or contradiction: The video does not relate to the claim at all, or it directly shows something opposite.
Based on the provided video and text, evaluate the probability that the text is true.
Your answer must be a decimal number between 0 and 1, and you must strictly follow the format below:
<answer>probability_value</answer>
Where probability_value is the result you calculate.
The text to evaluate is: {text}

Figure 6: Zero-Shot scaler prompt for vision examples.

```
To help you make more accurate and consistent judgments, here is an expanded explanation of how to
interpret and assign support percentages based on textual evidence. These examples are designed to
cover a range of logic and linguistic relationships you may encounter.
100% - Full and unambiguous support (Entailment): The sentence explicitly states the information in
the claim, or the claim is a direct paraphrase of the sentence. There is no need for guessing; the
facts are identical.
80-100% - Almost complete support: The main assertions in the claim are present in the sentence.
There may be minor differences in wording, synonyms, or omission of non-essential details, but the
core meaning is fully preserved and supported.
60-80% - Strong partial support (Strong Implication): The sentence strongly implies the claim is true
through logical inference or context, though it may not state it explicitly. A reasonable person
would conclude the claim is likely true based on the sentence.
40-60% - Moderate partial support: There is a topical alignment or shared keywords. The sentence
discusses the same subject matter, but the specific assertion in the claim is neither confirmed nor
denied. It is plausible but lacks definitive evidence in the text.
20-40% - Minimal weak support: There are weak textual links, such as matching entity names or a
general theme, but the specific context is different. The sentence provides very little basis to
deduce the claim.
0-20% - Very weak or speculative support: There may be a very distant connection (e.g., related
vocabulary), but inferring the claim from the sentence would be highly speculative.
0% - No support or contradiction: The sentence is completely unrelated to the claim, or it directly
contradicts the claim (proves it false).
Based on the provided sentence and claim, evaluate the probability that the claim is supported by
the sentence.
Your answer must be a decimal number between 0 and 1, and you must strictly follow the format below:
<answer>probability_value</answer>
Where probability_value is the result you calculate.
Sentence: {sentence}
Claim: {claim}
```

Figure 7: Zero-Shot scaler prompt for text examples.

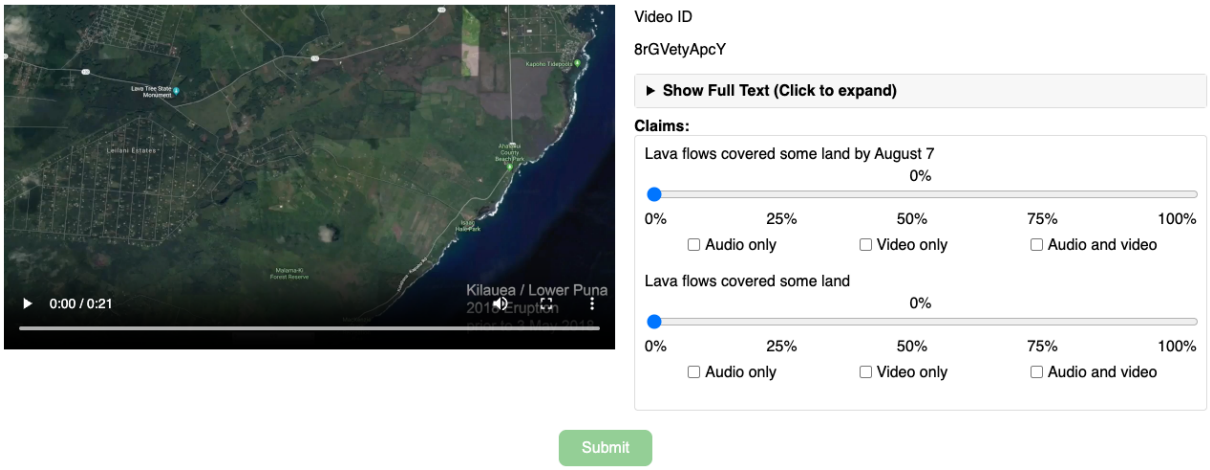

Figure 8: Annotation Protocol. Annotators are given a video (left) and set of claims (right) and can annotate the probability the claim is true with the supporting modalities.

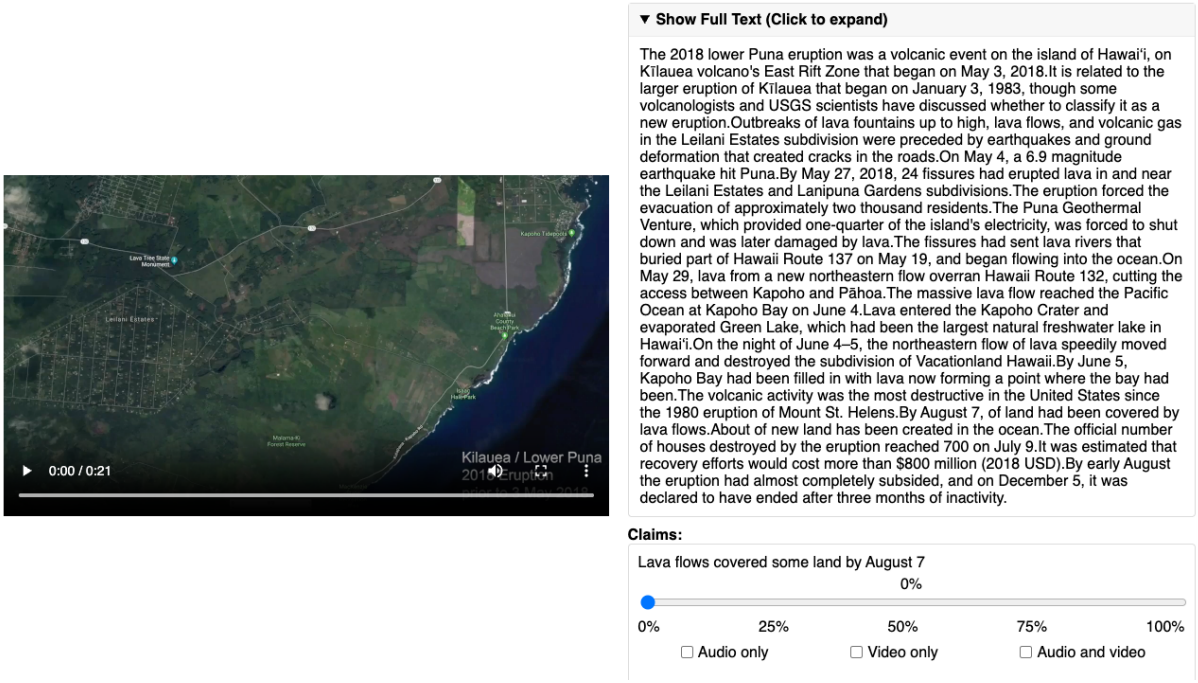

Figure 9: Annotation Protocol. Annotators can expand the annotation view to see the context from which the claim is taken to decontextualize and disambiguate any part of the claim.

Fine-grained Video Annotation

YOU MUST READ ALL INSTRUCTIONS BELOW BEFORE COMPLETING ANY HITS

Your goal in this task is to evaluate how likely a given claim is to be true based on a given video. You will be provided with the background story of an event, a video, and a series of claims related to the video. Your job is to determine the probability that each individual claim is true given the video, using a percentage scale.
When annotating you are allowed to use your background knowledge, but see the Background Story section for details on the type of knowledge you can use. However, for each event, you should treat it as it is currently happening (as depicted by the video) and not an event that has previously occurred. We suggest following the principle: "you can google information to make you more confident in claims, but not that would directly confirm or deny the claim itself."
Please follow the guidelines below when making your judgments:
100% — Select this if the video makes the claim almost certainly true. There should be direct and unambiguous evidence in the video that confirms the claim without the need for interpretation or inference. You should be confident in the statement beyond a reasonable doubt.
0% — Choose this if the video is completely unrelated to the claim or contradicts it. This means that nothing in the video provides evidence that the claim is true.
0%–100% — If the video provides partial evidence for the claim, select a percentage that best reflects how likely the claim is to be true given the video. For this part, you can refer to task instructions for more details.

What You Will See:
A story that provides context about the overall event. The story is to help you understand the types of background knowledge you may have, and the things you are able to search for while doing the annotation.
A video that related to an event.
One or more claims, each of which you must evaluate individually based on what is shown in the video.

Background Story and Knowledge
In order to ensure that annotators make accurate and consistent judgments when evaluating the probability that a claim is true given a video, we provide a short background story for each event. The story provides contextual information that helps annotators better understand the background knowledge (and searchable knowledge) an annotator may use while annotating the task.
Without such a story, annotators may rely solely on their own assumptions or incomplete information, which can lead to inconsistent or biased annotations. By giving all annotators the same background knowledge, we aim to reduce variability in the annotation process and encourage judgments that follow a more consistent distribution. For examples of how the presence or absence of a story can affect annotation quality, please refer to the "Examples" section.
We suggest following the principle: "you can google information to make you more confident in claims, but not that would directly confirm or deny the claim itself." To understand this better, have a look at the examples below.

Example 1, Notre-Dame Fire
Story: You're someone who regularly follows both local and international news. You're familiar with major cultural landmarks and global events. You are a resident of the location the event occurs in. If you are lacking background knowledge you are allowed to search for: maps of Paris and France, general knowledge a resident in France might have, etc.
Things you can google based on this event: What is a cathedral? What is gothic architecture?
Things you cannot google: How much donations were collected to restore the cathedral? When did the Notre-Dame start holding Christmas mass?

Example 2, 2022 Senate Election in Georgia:
Story: You are a resident living in Georgia and was going to watch a series of videos about the ongoing election. If you are lacking background knowledge you are allowed to search for: famous people, landmarks, local maps, and basic information about US elections.
Things you can google based on this event: When do elections occur in the US?
Things you cannot google based on this event: When did the 2022 senate election happen? Who won the 2022 senate election? Does Georgia have a two round election system?

Example 3, Hurricane Irma
Story: You are a Florida resident watching videos on news about this hurricane. If you are lacking background knowledge you are allowed to search for: maps labeling countries states, information to help you understand characteristics of hurricanes (within reason).
Things you can google based on this event: Map of Caribbean
Things you cannot google: Map of Caribbean showing the Hurricane Irma path

Task Instructions
Your task involves the following steps:
1. Read the background story to understand the context.
2. Watch the entire video clip carefully. Rewatch as many times as needed to fully absorb both visual and audio details. Sometimes you can refer to suggestions in the background story to get extra information.
3. Evaluate each claim on its own, using only the evidence presented within the video itself — including visuals, dialogue, audio cues, and on-screen text.
4. Assign a percentage score (0%–100%) to reflect how likely the claim is to be true given the video.

To help you make more accurate and consistent judgments, here is an expanded explanation of how to interpret and assign probability percentages. These examples are designed to cover a range of real-world cases you may encounter in the annotation task.
100% – Almost certainly true: The video clearly shows the exact event described in the claim. There is no need for guessing or interpretation.
80–100% – Very likely true: The main content in the claim is shown, but there may be minor ambiguity in location, identity, or completeness. The video makes the claim very likely to be true.
60–80% – Likely true: The video strongly suggests the claim is true, but some critical details may be missing, obscured, or ambiguous, limiting the ability to confirm the claim with certainty. The video provides strong but not definitive evidence.
40–60% – Uncertain: There is some alignment with the claim, but large portions are either missing, unclear, or open to interpretation. While the footage may point in the same general direction as the claim, it lacks the clarity or completeness needed for confident verification.
20–40% – Unlikely true: There are small visual or audio cues that could hint at the claim, but they are insufficient to judge the claim as likely true.
0–20% – Very unlikely true: There may be the slightest indirect reference, such as a related object or setting, but nothing concrete happens.
0% – Almost certainly false: The video does not relate to the claim at all, or it directly shows something opposite.
These ranges are intended to help you gauge how much evidence is present and how likely the claim is to be true given that evidence. When in doubt, lean toward the lower end of the range unless you have clear reason to be confident. Remember: the key is not whether the claim is plausible in general — only whether the video provides evidence that it is true.

What Does It Mean to Estimate the Probability of a Claim?
You can think of each claim as posing a yes-or-no question: is this claim true based on the video? Your task is to estimate the probability that the answer is yes. If the video (through its visuals, audio, or on-screen text) provides evidence that the claim is true, the probability should be higher.
Evidence can come from:
Audio: Speech or sounds that indicate the claim is true.
Visual: Actions, events, or imagery shown on-screen that verify the claim.
Text: On-screen text such as subtitles, signs, or labels that confirm the claim.
If you meet something you don't understand or you are confused about, you can get some information in the story which would help. If the video does not provide any evidence relevant to the claim, or even contradicts it, then the probability should be low (near 0%).

Figure 10: Annotation instructions provided to annotators for scalar probability judgments. Along with these instructions, annotators are also provided a set of example annotations done by authors of the paper.