Adaptive Evidential Learning for Temporal-Semantic Robustness in Moment Retrieval

Haojian Huang[1,2*], Kaijing Ma[2,3*], Jin Chen[2,3,5*], Haodong Chen[1], Zhou Wu[6], Xianghao Zang[2], Han Fang[2], Chao Ban[2], Hao Sun[2†], Mulin Chen[4,2†], Zhongjiang He[2]

[1]Hong Kong University of Science and Technology (Guangzhou),
[2]Institute of Artificial Intelligence (TeleAI), China Telecom
[3]Fudan University,
[4]Northwestern Polytechnical University,
[5]Shanghai Innovation Institute
[6]Independent Researcher

hhuang118@connect.hkust-gz.edu.cn, chenmulin001@gmail.com, sun.010@163.com

Abstract

In the domain of moment retrieval, accurately identifying temporal segments within videos based on natural language queries remains challenging. Traditional methods often employ pre-trained models that struggle with fine-grained information and deterministic reasoning, leading to difficulties in aligning with complex or ambiguous moments. To overcome these limitations, we explore Deep Evidential Regression (DER) to construct a vanilla Evidential baseline. However, this approach encounters two major issues: the inability to effectively handle modality imbalance and the structural differences in DER's heuristic uncertainty regularizer, which adversely affect uncertainty estimation. This misalignment results in high uncertainty being incorrectly associated with accurate samples rather than challenging ones. Our observations indicate that existing methods lack the adaptability required for complex video scenarios. In response, we propose Debiased Evidential Learning for Moment Retrieval (**DEMR**), a novel framework that incorporates a Reflective Flipped Fusion (RFF) block for cross-modal alignment and a query reconstruction task to enhance text sensitivity, thereby reducing bias in uncertainty estimation. Additionally, we introduce a Geom-regularizer to refine uncertainty predictions, enabling adaptive alignment with difficult moments and improving retrieval accuracy. Extensive testing on standard datasets and debiased datasets ActivityNet-CD and Charades-CD demonstrates significant enhancements in effectiveness, robustness, and interpretability, positioning our approach as a promising solution for temporal-semantic robustness in moment retrieval.

Code — https://github.com/KaijingOfficial/DEMR

Introduction

Moment Retrieval (MR) in video understanding is a pivotal task that involves locating specific time segments within untrimmed videos based on natural language queries (Hu et al. 2024). This task demands models to effectively integrate and interpret both visual and textual information to accurately identify relevant moments. Traditional approaches (Lin et al. 2023; Liu et al. 2024a; Lei, Berg, and Bansal 2021a; Li et al. 2024; Moon et al. 2023b; Huang et al. 2022) predominantly rely on pretrained Transformer models such as CLIP-ViT (Radford et al. 2021a), which are primarily pretrained on instance-level classification, thereby limiting their ability to address the nuanced demands of MR in complex video scenarios. Even when models enhance their ability to align visual and textual features, they often operate under determin-

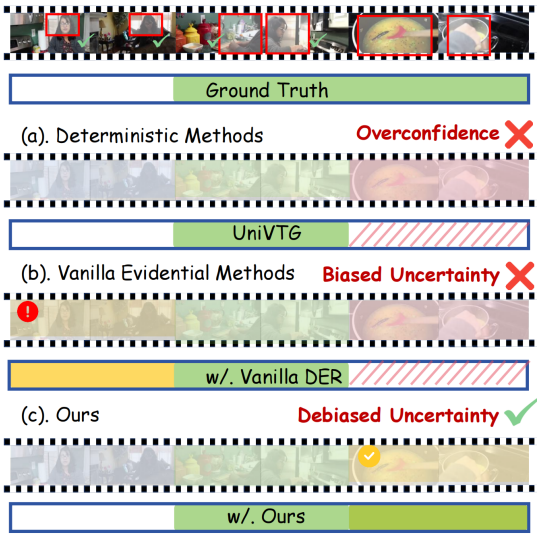

Figure 1: Comparison of MR methods: (a) Deterministic methods (*e.g.* (Lin et al. 2023)) are overconfident with limited evidence, using Non-Maximum Suppression (NMS) but still failing on challenging frames; (b) Vanilla evidential methods consider uncertainty but produce biased estimates on hard samples; (c) Our method adaptively aligns with challenging semantics for accurate uncertainty modeling and improved inference. Yellow regions denotes uncertainty predictions.

*Equal Contribution
†Corresponding Author

istic paradigms. These methods struggle to provide accurate predictions in challenging video frames. As illustrated in Figure 1, when frames lack the presence of a woman, models find it difficult to align with the query "cooking." And during inference, techniques like Non-Maximum Suppression (NMS) are used to select the most probable segments. But without additional knowledge, these methods fail to align with hard moments. To address this, we integrate Deep Evidential Regression (DER) (Amini et al. 2020) into a vanilla evidential method as a baseline (*i.e.* Figure 1(b)). Based on Evidential Deep Learning (EDL) (Sensoy, Kaplan, and Kandemir 2018), DER excels at capturing uncertainty by learning a second-order distribution to fit the correct moment proposals. Unlike deterministic methods, DER represents uncertainty by treating each proposal as evidence sampled from observations. This approach learns the correspondence between evidence and ground truth, aligning a second-order distribution to the target domain. Through this learned distribution, the model assesses sample uncertainty and adjusts gradient backpropagation, resulting in more robust and accurate inference. This capability has already led to significant success in various regression tasks (Wang et al. 2022a; Wu, Zhang, and Woodland 2023; Ye et al. 2024; Ma et al. 2024).

However, applying DER directly to MR presents challenges. Unlike other unimodal domains (Wu, Zhang, and Woodland 2023; Ye et al. 2024), evidence requires comprehensive fusion of visual and textual information in MR. Simple concatenation of multimodal features fails to ensure nuanced understanding. Additionally, DER faces counterintuitive uncertainty issues: higher-error predictions can receive lower uncertainty due to limitations in vanilla DER's regularizer. Unlike classification-based EDL methods, DER lacks a standard KL-divergence term, relying on a heuristic regularizer that overly suppresses evidence, especially in low-error samples, misaligning uncertainty estimates with low-error samples showing higher uncertainty and vice versa.

To address modality imbalance, we propose Debiased Evidential Learning for Moment Retrieval (**DEMR**). Our approach incorporates a Reflective Flipped Fusion (RFF) block with dual branches for progressive cross-modal alignment, complemented by a query reconstruction (QR) task to strengthen the text branch. This design enhances the model's sensitivity to textual information and mitigates bias in uncertainty estimation. To resolve counterintuitive uncertainty, we introduce a simple yet effective Geom-regularizer, which adjusts uncertainty estimation based on prediction accuracy, adaptively suppressing overconfidence and debiasing the system. Our contributions are summarized as follows:

- We introduce an uncertainty-aware MR baseline with DER, and further develop DEMR to address challenging and ambiguous moments.
- Our approach integrates the RFF block, auxiliary QR task, and Geom-regularizer to mitigate modality imbalance and improve uncertainty estimation, enabling adaptive alignment with difficult moments.
- Experiments on common datasets and debiased versions of ActivityNet and Charades-STA, demonstrate effectiveness, robustness and interpretability.

Related Work

Moment Retrieval focuses on localizing temporal segments in untrimmed videos based on textual queries, and has evolved rapidly in recent years. Early MR methods adopted two-stage frameworks (Zhang, Su, and Luo 2019; Zhang et al. 2019c; Gao and Xu 2021), generating temporal proposals followed by refinement steps such as Non-Maximum Suppression (NMS), but these approaches were computationally expensive and relied heavily on hand-crafted priors. One-stage models (Chen et al. 2018; Wang, Ma, and Jiang 2020; Otani et al. 2020a; Zhang et al. 2019a; Hu et al. 2021; Liu et al. 2018; Zhang et al. 2020) aimed to directly predict moment boundaries, improving efficiency but still faced challenges in flexibility and alignment. The introduction of transformer-based architectures, notably Detection Transformer (DETR) (Carion et al. 2020), reframed MR as a set prediction problem, inspiring models such as Moment-DETR (Lei, Berg, and Bansal 2021a), QD-DETR (Moon et al. 2023b), and MESM (Liu et al. 2024c), which further advanced cross-modal alignment and prediction accuracy. However, despite these advances, most state-of-the-art MR methods remain deterministic and lack effective mechanisms for modeling uncertainty, limiting their robustness in complex or ambiguous scenarios.

Recently, uncertainty Learning has gained attention as MR datasets are found to contain inherent ambiguities and biases (Zhang et al. 2023). Annotation uncertainty arises from inconsistent temporal boundaries set by different annotators, while query uncertainty stems from diverse natural language descriptions for the same video moment. Additionally, dataset bias—such as overrepresentation of common events and long-tailed distributions—further challenges model generalization (Zhang et al. 2023; Otani et al. 2020b). These issues highlight the need for uncertainty-aware modeling. Evidential Deep Learning (EDL), grounded in Dempster-Shafer Theory (Shafer 1992) and Subjective Logic (Sensoy, Kaplan, and Kandemir 2018; Jøsang 2016), explicitly models uncertainty via second-order probability distributions and has shown effectiveness in classification (Chen, Huang, and Li 2024; Huang et al. 2024b; Holmquist, Klasén, and Felsberg 2023; Liu et al. 2024b; Huang et al. 2024a, 2025) and regression tasks (Amini et al. 2020; Wang et al. 2022a; Wu, Zhang, and Woodland 2023). Deep Evidential Regression (DER) extends EDL to regression, but suffers from evidence contraction and gradient issues under high uncertainty (Wu et al. 2024; Ye et al. 2024). Recent advances have introduced new regularizers to address these limitations, but their application to MR remains unexplored. In this work, we bridge this gap by introducing DER into the MR task, aiming to improve model robustness and reliability under uncertainty. We further address modality imbalance and structural limitations of vanilla DER, marking the first successful extension of evidential regression to moment retrieval.

Preliminaries

DER (Amini et al. 2020) places evidential priors over the original Gaussian likelihood function and trains the model to infer the hyperparameters of the evidential distribution. This

approach enables the model to learn both aleatoric and epistemic uncertainty. In our context, adjacent video frames often exhibit similar semantics, which introduces uncertainty in precisely locating temporal boundaries. The start or end temporal boundary of a video is represented by distinct Gaussian distributions: $\boldsymbol{b} \sim \mathcal{N}(\mu, \sigma^2)$, where $\boldsymbol{b} \in \mathbb{R}^{1\times\mathcal{H}}$ represents the start or end of moments observed $\mathcal{H}$ times. We assume that observations of the same type (either all starts or all ends) are *i.i.d.*. The corresponding expectation μ and variance σ^2 of the Gaussian distribution subject to NIG prior:

$$p(\mu, \sigma^2 \mid \underbrace{\gamma, \upsilon, \alpha, \beta}_{\boldsymbol{\varphi}}) = \mathcal{N}(\mu|\gamma, \sigma^2\upsilon^{-1})\Gamma^{-1}(\sigma^2|\alpha, \beta), \quad (1)$$

where $\boldsymbol{\varphi} = (\gamma, \upsilon, \alpha, \beta)$ are the prior NIG distribution parameters derived from the video content and user queries, serve as conditionals for the Gaussian estimates of b_i with $\gamma \in \mathbb{R}, \upsilon > 0, \alpha > 1, \beta > 0$. The gamma function is denoted by $\Gamma(\cdot)$. We use a linear evidential predictor to estimate $\boldsymbol{\varphi}$, training it to maximize the likelihood. Since the likelihood function has a form of Student-t distribution (St), we minimize the negative logarithmic likelihood (NLL) as follows:

$$\mathcal{L}_i^{\mathrm{NLL}} = -\log p(b_i|\boldsymbol{\varphi}) = -\log\left(\mathrm{St}\left(b_i; \gamma, \frac{\beta(1+\upsilon)}{\upsilon\alpha}, 2\alpha\right)\right), \quad (2)$$

Models optimized only on observed samples with the NLL loss (*i.e.* Eq. 2) tend to overfit and exhibit overconfidence. To counter this, DER introduced a regularizer for the i-th prediction as follows:

$$\mathcal{L}_i^{\mathrm{R}}(\boldsymbol{\vartheta}) = \Delta \cdot \Phi, \quad (3)$$

where $\Delta = |b_i - \gamma|$ represents the error, $\Phi = 2\upsilon + \alpha$ denotes the evidence, and $\boldsymbol{\vartheta}$ are the model parameters, with b_i as the ground truth. Detailed formulation can be found in **Supplementary Material**. Using the NIG distribution, prediction, aleatoric and epistemic uncertainties are calculated as follows:

$$\underbrace{\mathbb{E}[\mu] = \gamma}_{\text{prediction}}, \quad \underbrace{\mathbb{E}[\sigma^2] = \frac{\beta}{\alpha - 1}}_{\text{aleatoric}}, \quad \underbrace{\mathrm{Var}[\mu] = \frac{\beta}{\upsilon(\alpha-1)}}_{\text{epistemic}}, \quad (4)$$

$\mathbb{E}[\sigma^2]$ refers to the inherent noise in the data, which cannot be reduced or eliminated. $\mathrm{Var}[\mu]$ reflects the model's lack of confidence in its own predictions due to limited knowledge.

Methodology

Problem Definition

Given a video $V = \{v_i\}_{i=1}^{L_v}$ and a query $Q = \{q_i\}_{i=1}^{L_q}$, both as vectors in $\mathbb{R}^D$, Moment Retrieval aims to locate the time span $m = [m^s, m^e]$ in the video that best matches Q, by identifying the start and end clips m^s and m^e of the relevant segment.

Building Baseline with Vanilla DER for MR

Deterministic methods in moment retrieval struggle with nuanced video frames and often fail to align with challenging moments due to their reliance on instance-level classification. In contrast, DER captures uncertainty by learning a second-order distribution, treating each proposal as evidence. This approach enhances robustness and accuracy, allowing for better alignment with complex video scenarios. Therefore, as illustrated in Figure 2 (a), we first build an uncertainty-aware baseline by integrating DER into the MR task. The motivation for this is to tackle the challenging video frames in moment retrieval, enabling the model to progressively align difficult samples semantically through uncertainty representation. Overall, the loss function of the model can be formulated as follows:

$$\mathcal{L}_i^{\mathrm{B}}(\boldsymbol{w}) = \lambda_{\mathrm{NLL}}\mathcal{L}_i^{\mathrm{NLL}} + \lambda_{\mathrm{Reg}}\mathcal{L}_i^{\mathrm{R}}(\boldsymbol{w}), \quad (5)$$

$$\mathcal{L}_{base} = \mathcal{L}_{\mathrm{mr}} + \lambda_{\mathrm{der}}\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_i^{\mathrm{B}}(\boldsymbol{w}), \quad (6)$$

where N symbolizes the number of clips in a training set and $\mathcal{L}_{\mathrm{mr}}$ denotes MR loss (*i.e.* Eq. 10). While DER effectively estimates uncertainty, its vanilla form presents limitations like modality imbalance and flawed uncertainty estimation, which our baseline exposes and sets the foundation for improvement.

Debiased DER Model for MR

To address the biased uncertainty estimation in the baseline in section , caused by modality imbalance and counterintuitive uncertainty, we propose **DEMR**. Our model introduces a RFF block for progressive cross-modal alignment, reducing over-reliance on visual features, and a QR task to enhance text sensitivity. As shown in Figure 2 (b), DEMR first encodes an untrimmed video and masked query, reconstructs the masked tokens via the RFF block, and performs MR. The debiased DER head assesses both aleatoric and epistemic uncertainties, while the MR and QR heads manage task stages. Further details are provided in subsequent sections.

RFF Block. The RFF block iteratively updates video and text features by alternating their roles as queries and keys/-values in a shared cross-attention (CA) module, followed by self-attention (SA) refinement in each branch. At each layer, features are updated as:

$$V^{(i+1)} = SA_v^{(i)}(CA_{q\to v}^{(i)}), \quad Q^{(i+1)} = SA_q^{(i)}(CA_{v\to q}^{(i)}) \quad (7)$$

where CA and SA are defined as:

$$CA_{v\to q}^{(i)} = \text{Softmax}\left(\frac{V^{(i)}Q^{(i)\mathsf{T}}}{\sqrt{d_k}}\right)Q^{(i)} \quad (8)$$

$$SA_v^{(i)} = \text{Softmax}\left(\frac{CA_{v\to q}^{(i)}CA_{v\to q}^{(i)\mathsf{T}}}{\sqrt{d_k}}\right)CA_{v\to q}^{(i)} \quad (9)$$

This process repeats for n layers, progressively enhancing cross-modal alignment.

MR Head. Given $\tilde{\mathbf{V}}^k \in \mathbb{R}^{L_v\times D}$, this head generates a series of offsets $\{\tilde{m}_i\}_{i=1}^{L_v}$ for each unit. We then define the predicted boundary $\tilde{m}_i$ and the corresponding interval d_i (i.e., $d_i = m_i^s - m_i^e$). For training objectives, we use a combination

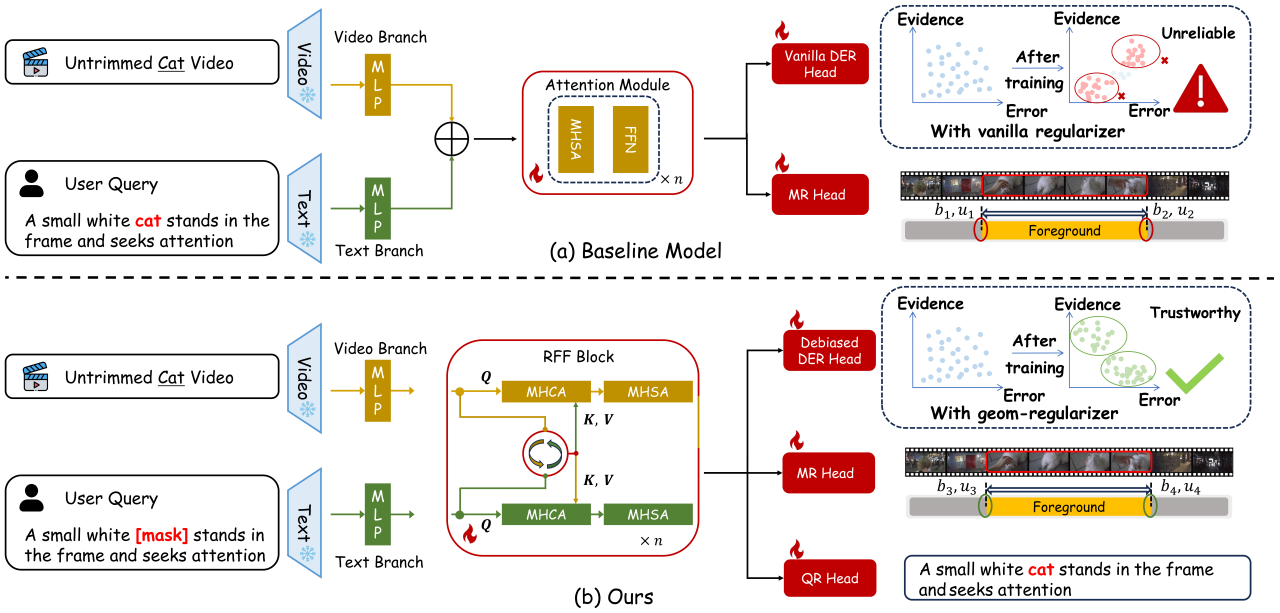

Figure 2: Comparison of the baseline (a) and our improved model (b) for the MR task. In (a), the baseline exhibits weak sensitivity to text, as the overlap between the MR task and DER objective causes over-reliance on visual features, while the vanilla DER regularizer leads to unreliable uncertainty estimates. In (b), our RFF block and QR head enhance cross-modal interaction and text sensitivity, and the Geom-regularizer corrects structural flaws in DER for more reliable uncertainty estimation.

of smooth L1 loss and generalized IoU loss to optimize the model's performance.

$$\mathcal{L}_{\text{mr}} = \mathbb{1}_{f_i=1}\left[\lambda_{\text{L1}}\mathcal{L}_{\text{SmoothL1}}\left(\tilde{d}_i, d_i\right) + \lambda_{\text{iou}}\mathcal{L}_{\text{iou}}\left(\tilde{m}_i, m_i\right)\right]. \tag{10}$$

Notably, this regression objective is only devised for foreground clips *i.e.*, $f_i = 1$.

Query Reconstruction Task. We observe that nouns are crucial for cross-modal tasks, as CLIP features primarily capture objects due to training on static images, leading to a focus on static visual cues rather than dynamic actions. In our benchmarks, identifying key nouns is often sufficient for high-quality reasoning, motivating our QR task design. To enhance cross-modal alignment, we mask query entities at a fixed ratio during early alignment, forcing the model to employ both video context and remaining query tokens. The QR head then reconstructs the masked tokens, with a dedicated loss optimizing cross-modal inference as follows:

$$\mathcal{L}_{qr} = \mathbb{E}\left[-\sum_{i=1}^{l}\log P(w_i \mid U, V)\right], \tag{11}$$

Here, l is the number of masked tokens, w_i denotes the i-th masked token, U the unmasked query tokens, and V the video features supporting accurate prediction. After the initial alignment phase, the QR head is frozen and $\mathcal{L}_{qr}$ is excluded from training and inference.

Geom-Regularization. The heuristic regularizor (*i.e.* Eq.(3)) in conventional DER aims to mitigate overconfidence by suppressing evidence, particularly for samples with high

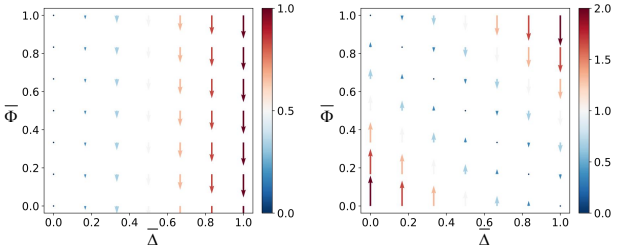

Figure 3: Gradient field comparison. (a) Vanilla regularizer applies penalties based solely on error, decreasing evidence as error increases. (b) Our Geom-regularizer modulates penalties dynamically based on error magnitude and evidence levels. Our approach reflects the principle that accurate predictions should have higher evidence, while evidence should be suppressed for less accurate predictions.

error. However, excessive suppression can lead to underconfidence due to non-adaptive suppression and sample imbalance. To be clear, we first consider the minus gradient of $\mathcal{L}_i^{\text{R}}$ for Φ as follows:

$$-\nabla_{\Phi}\mathcal{L}_i^{\text{R}} = -\Delta, \tag{12}$$

To explore the penalties bias in the vanilla regularizer, we visualized its optimization direction by examining the gradient field derived from the Eq.(12). As shown in Fig 3 (a), the gradient is solely linked to the error and not to the evidence, indicating that the model cannot ascertain when the evidence has been adequately suppressed. This approach often results in insufficient gradients for batches dominated by small er-

rors, potentially leading to biased penalties on evidence. As the model converges, the dominance of low error samples with small gradients skews the batch's average gradient. Consequently, their evidence is over-suppressed, while high error samples see their evidence neglected or adversely adjusted, as shown in Figure 5. To overcome these limitations, we introduce Geom-regularization, inspired by (Amini et al. 2020), promoting the principle that "***accurate predictions should have high evidence, while inaccurate ones should have low evidence***". This approach provides more rational constraints rather than merely suppressing evidence. Initially, we normalize Δ to $\overline{\Delta}$ and Φ to $\overline{\Phi}$ (details in **supplementary material**), which ensures that the model assigns $\overline{\Phi} = 1$ to samples with $\overline{\Delta} = 0$, and $\overline{\Phi} = 0$ to samples with $\overline{\Delta} = 1$. We then ensure that the points $(\overline{\Delta}, \overline{\Phi})$ closely follow the line $\overline{\Phi} + \overline{\Delta} = 1$ using a line regularizer as below:

$$\mathcal{L}_i^{\mathrm{L}}(\boldsymbol{w}) = \|\overline{\Phi} + \overline{\Delta} - 1\|_2^2, \tag{13}$$

we can follow the analysis for $\mathcal{L}_i^{\mathrm{R}}$. The minus gradient of $\mathcal{L}_i^{\mathrm{L}}$ with respect to $\overline{\Phi}$ as below:

$$-\nabla_{\overline{\Phi}} \mathcal{L}_i^{\mathrm{L}} = -2(\overline{\Delta} + \overline{\Phi} - 1), \tag{14}$$

which indicates this simple regularizer offers a gradient that relates to both error and evidence, enabling adaptive evidence suppression, as illustrated in Figure 3 (b).

Our training objective for the evidential head is the combination of NLL and Geom-regularization:

$$\mathcal{L}_i^{\mathrm{e}}(\boldsymbol{w}) = \lambda_{\mathrm{NLL}} \mathcal{L}_i^{\mathrm{NLL}} + \lambda_{\mathrm{geom}} \mathcal{L}_i^{\mathrm{L}}(\boldsymbol{w}), \tag{15}$$

To this end, our total loss can be formulated by a combination of MR loss $\mathcal{L}_{\mathrm{mr}}$ and our evidential loss:

$$\mathcal{L} = \mathcal{L}_{\mathrm{mr}} + \lambda_{\mathrm{der}} \frac{2}{N} \sum_{i=1}^{N} \mathcal{L}_i^{\mathrm{e}}(\boldsymbol{w}) + \mathcal{L}_{qr}, \tag{16}$$

where N symbolizes the number of clips in a training set.

Experiment

Datasets and Implementation Details

Datasets and Metrics. We evaluate on diverse public datasets: Charades-STA (Gao et al. 2017) (indoor activities), QVHighlights (Lei, Berg, and Bansal 2021c) (untrimmed vlogs/news), and TACoS (Regneri et al. 2013) (cooking scenes). To assess robustness under temporal bias, we also use debiased ActivityNet-CD and Charades-CD (Lan et al. 2022). Dataset details and hyperparameters are provided in the **supplementary material**. And we report Recall@1 at IoU 0.5/0.7, mAP@0.75 and mAP avg (mean MAP over IoU 0.5-0.95, step 0.05), namely mAP.

Experimental Settings. Following (Lin et al. 2023; Li et al. 2024), we use CLIP (Radford et al. 2021b) (ViT-B/32) and SlowFast (Feichtenhofer et al. 2019) (ResNet-50) as frozen backbones. The number of RFF blocks is set to 4. Training is two-stage: (1) QR masks/reconstructs 1 noun per sentence (default), using spaCy (Honnibal and Montani 2017) for noun extraction; QR runs for 30 epochs with a 1e-5 learning rate. (2) DEMR predicts bounding boxes per video clip; DER gradients in Eq. (13) are detached to focus on uncertainty optimization. NMS with threshold 0.7 is applied at evaluation. Unless stated, line regularizer is used on the evidential head. All experiments run on four Tesla V100 GPUs.

Quantitative Results

Comparison with the state-of-the-art. For fair evaluation, we adopt the same backbone as most compared methods in both Table 1 and Table 2. DEMR is benchmarked against leading traditional and MLLM-based approaches, demonstrating notable competitiveness across all datasets. Although this work primarily focuses on robust evidential learning for MR tasks, DEMR still achieves strong results without relying on dense clip-word guidance (CG-DETR) or dense frame sampling and temporal encoding (LLaVA-MR). This highlights the effectiveness of our approach and suggests substantial potential for further improvement when integrated with advanced MLLM-based backbones.

Ablation Study. To evaluate the effectiveness of our debiasing method, we introduce targeted metrics to quantify modality imbalance on the QVHighlights validation set (Lei, Berg, and Bansal 2021c). Specifically, we define Varvis, $\mathrm{Vartext}$, and Δ_{Var} to measure uncertainty sensitivity under controlled noise. Gaussian noise is added to video embeddings, and a proportion of text tokens is replaced with irrelevant content. We compute the average variance of uncertainty for each modality and use Δ_{Var} to assess balance. As shown in Table 3(a), both the RFF block and QR task substantially reduce modality bias, confirming the effectiveness of our approach.

Qualitive Resluts

Uncertainty Calibration. We assess the calibration efficacy on QVHighlights in Figure 4. Without proper regularization (w/o DER or NLL-only), the model tends to be overconfident, ignoring errors. The vanilla regularizer (Amini et al. 2020) presents a paradox by showing high uncertainty even at low error rates. Conversely, our Geom-Regularizer effectively corrects these issues by ensuring uncertainty grows monotonically with the normalized error $\overline{\Delta}$, thereby achieving reliable uncertainty estimation.

As illustrated in Figure 5, it is obvious that "***accurate predictions with high evidence while inaccurate predictions with low evidence***" has been reflected in the knowledge of model with only NLL. Unfortunately, the vanilla regularizer excessively suppress the evidence of low error predictions, but ignores and even enlarges the evidence of high error predictions. Geom-regularizer turn the situation around, retain the main knowledge learned by NLL, and provides calibration for more reasonable uncertainty estimation.

Temporal Bias Sensitivity. Most moment retrieval datasets exhibit significant temporal bias in moment duration and position, leading to underrepresented (Temporal OOD) regions, as visualized in Figure 4(a) using QVHighlights. Higher epistemic uncertainty is expected in these OOD regions. We evaluate model uncertainty across temporal regions under different settings. Without DER constraints, the evidential head's uncertainty merely reflects the biased data distribution (Figure 4(b)). Using only NLL loss, the model shows

Method	QVHighlights			TACoS			Charades-STA		
	R1@0.5	R1@0.7	mAP	R1@0.5	R1@0.7	mIoU	R1@0.5	R1@0.7	mIoU
M-DETR (Lei, Berg, and Bansal 2021b)	53.9	34.8	30.7	28.0	12.9	27.2	46.0	27.5	41.3
UMT (Liu et al. 2022)	60.3	44.3	38.6	23.5	13.2	25.0	42.7	24.1	41.6
QD-DETR (Moon et al. 2023c)	62.7	46.7	41.2	24.7	12.0	25.5	52.1	30.6	45.5
UniVTG (Lin et al. 2023)	59.7	-	36.1	35.0	17.4	33.6	58.0	35.7	50.1
CG-DETR (Moon et al. 2023a)	**67.4**	**52.1**	42.9	**39.5**	**23.4**	**37.4**	58.4	36.3	50.1
MomentDiff (Li et al. 2024)	57.4	39.7	36.0	33.7	-	-	55.6	32.4	-
DEMR (Ours)	65.0	49.4	**43.0**	37.3	19.4	33.9	**60.2**	**38.0**	**51.6**

Table 1: Performance comparison with traditional ViT- and CNN-based methods on QVHighlights (*val.* set), TACoS, and Charades-STA. Bold: best, underline: second. This table's results are not based on any additional pre-training data.

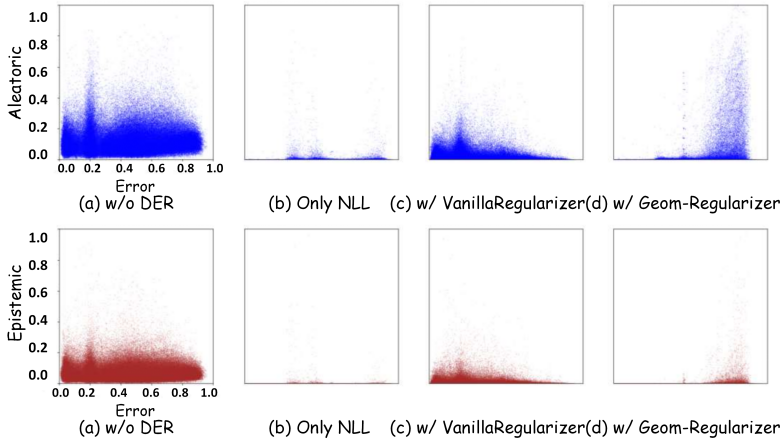

Figure 4: Effects of Various Regularization on Uncertainty Distribution. (a)-(d) illustrate the impact of different regularization on the relationship between aleatoric uncertainty (top row) and epistemic uncertainty (bottom row) with respect to prediction error. The models include: (a) without DER, (b) only NLL, (c) with Vanilla Regularizer, and (d) with Geom-Regularizer.

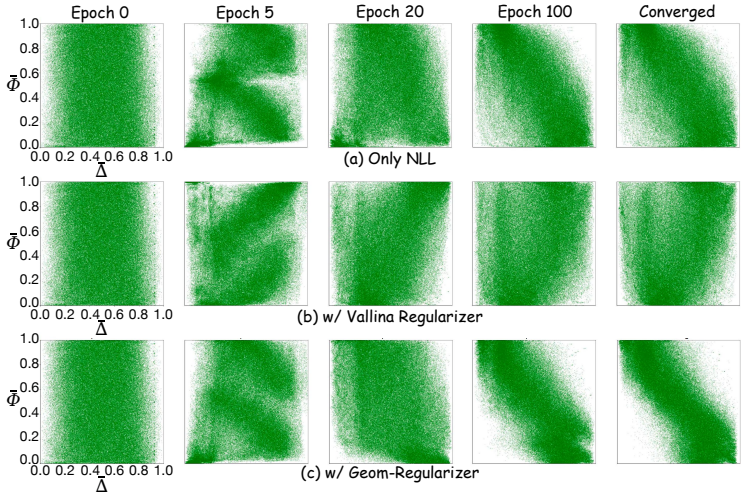

Figure 5: Illustration of the different regularizer impact. As the training progresses, the model's uncertainty is optimized in the right direction with our regularizer.

low epistemic uncertainty, indicating overconfidence (Figure 4(c)). Vanilla regularization reduces concentrated uncertainty but lacks OOD sensitivity (Figure 4(d)). In contrast, our Geom-regularizor significantly increases epistemic uncertainty in OOD regions (Figure 4(e)), enhancing temporal bias awareness. Table 4 and 5 further shows that DEMR achieves robust cross-distribution performance, notably surpassing MomentDiff and MomentDETR by 3.37% and 10.47% respectively on Charades-CD R1@0.3. Importantly, DEMR narrows the IID-OOD gap to 3.29%, compared to 12.00%

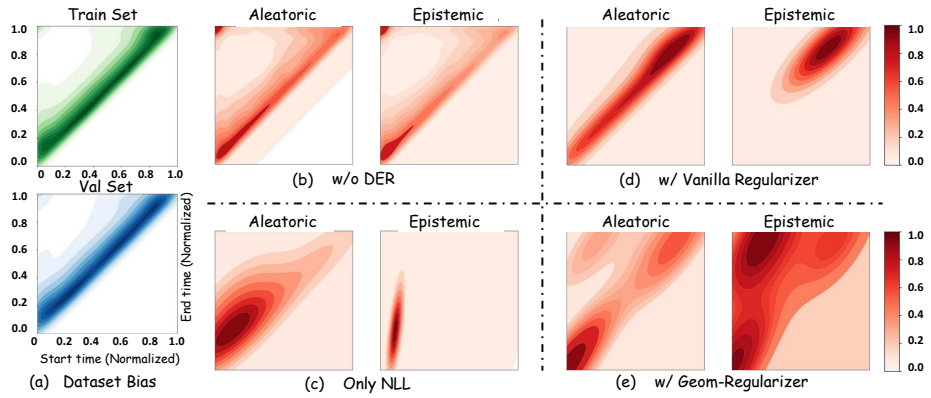

Figure 6: Dataset bias sensitivity. (a) Joint distributions of the start and end timestamps of the ground-truth moments in the QVHighlights dataset. (b), (c), (d), and (e) show the predicted uncertainty's sensitivity to bias in dataset under different settings.

Method	R1@0.5	R1@0.7	mAP@0.75	mAP
Mr. BLIP (Meinardus et al. 2024)	74.77	60.51	53.38	51.37
LLaVA-MR (Lu et al. 2024)	**76.59**	61.48	54.40	–
Ours	76.36	**62.91**	**56.82**	**52.32**

Table 2: Comparison on QVHighlights *test* set with prevalent methods based on Multimodal Large Language Model (MLLM) backbone, *i.e.* BLIP-2.

Method	R1@0.5	Var_{vis}	Var_{text}	$\Delta_{\text{Var}} \downarrow$
Baseline	61.1	9.17	0.85	8.32
+ RFF block	62.4	8.63	1.60	7.03
+ QR	63.8	4.89	3.91	0.98
Full Model	**65.0**	4.85	5.54	**0.69**

Table 3: Ablation studies on the QVHighlights validation split. (a) Component Ablation: Var_{vis} and Var_{text} denote uncertainty variance as noise is added to visual or textual inputs, reflecting DEMR's sensitivity to each modality.

for CM-NAT, demonstrating improved generalization and reduced sensitivity to dataset bias. These results confirm DEMR's effectiveness in addressing temporal bias and enhancing adaptability in challenging retrieval scenarios.

Conclusion

As AGI advances, MR models face increasing challenges from open-ended user inputs. In this paper, we propose DEMR, a robust MR model that addresses key biases in baseline evidential-based methods and enables effective uncertainty quantification, thus improving response credibility for hard cases. While current performance is constrained by data quality and scale, DEMR provides valuable strategies for enhancing the trustworthiness of AI decisions. Future work will further integrate DEMR with advanced MLLMs to expand its application and improve reliability in video tasks.

Method	**R1@0.3**	**R1@0.5**	**R1@0.7**
DRN (Zeng et al. 2020) (i.i.d)	51.35	41.91	26.74
DRN (Zeng et al. 2020) (o.o.d)	40.45	30.43	15.91
Δ (↓)	10.90	11.48	10.83
TSP-PRL (Wu et al. 2020) (i.i.d)	46.44	35.43	17.01
TSP-PRL (Wu et al. 2020) (o.o.d)	31.93	19.37	6.20
Δ (↓)	14.51	16.06	10.81
2D-TAN (Zhang et al. 2019b) (i.i.d)	53.71	46.48	28.18
2D-TAN (Zhang et al. 2019b) (o.o.d)	43.45	28.18	13.73
Δ (↓)	10.26	18.30	14.45
MMN (Wang et al. 2022b) (o.o.d)	55.91	34.56	15.84
CM-NAT (Lan et al. 2023) (i.i.d)	64.21	53.82	34.47
CM-NAT (Lan et al. 2023) (o.o.d)	52.21	39.86	21.38
Δ (↓)	12.00	13.96	13.09
MomentDETR (Lei, Berg, and Bansal 2021a)	57.34	41.18	19.31
MomentDiff (Li et al. 2024) (o.o.d)	67.73	47.17	22.98
Ours (i.i.d)	**71.10**	**62.20**	**43.29**
Ours (o.o.d)	**67.81**	**52.46**	**30.97**
Δ (↓)	**3.29**	**9.74**	**12.32**

Table 4: Performance comparison on Charades-CD.

Method	**R1@0.3**	**R1@0.5**	**R1@0.7**
DRN (Zeng et al. 2020) (i.i.d)	48.92	39.27	25.71
DRN (Zeng et al. 2020) (o.o.d)	36.86	25.15	14.33
Δ (↓)	12.06	14.12	11.38
TSP-PRL (Wu et al. 2020) (i.i.d)	44.93	33.93	19.50
TSP-PRL (Wu et al. 2020) (o.o.d)	29.61	16.63	7.43
Δ (↓)	15.32	17.30	12.07
2D-TAN (Zhang et al. 2019b) (i.i.d)	49.18	40.87	28.95
2D-TAN (Zhang et al. 2019b) (o.o.d)	30.86	18.86	9.77
Δ (↓)	18.32	22.01	19.18
MMN (Wang et al. 2022b) (o.o.d)	44.13	24.69	12.22
CM-NA (Lan et al. 2023) (i.i.d)	49.91	41.67	28.82
CM-NA (Lan et al. 2023) (o.o.d)	32.32	20.78	11.03
Δ (↓)	17.59	20.89	17.79
MomentDETR (Lei, Berg, and Bansal 2021a)	39.98	21.30	10.58
MomentDiff (Li et al. 2024) (o.o.d)	45.54	26.96	13.69
Ours (i.i.d)	**56.33**	**41.77**	**27.47**
Ours (o.o.d)	**41.64**	**23.76**	**16.89**
Δ (↓)	**14.69**	**18.01**	**10.58**

Table 5: Performance comparison on ActivityNet-CD.

References

Amini, A.; Schwarting, W.; Soleimany, A.; and Rus, D. 2020. Deep evidential regression. *Advances in neural information processing systems*, 33: 14927–14937.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Chen, J.; Chen, X.; Ma, L.; Jie, Z.; and Chua, T.-S. 2018. Temporally grounding natural sentence in video. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, 162–171.

Chen, M.; Huang, H.; and Li, Q. 2024. Towards Robust Uncertainty-Aware Incomplete Multi-View Classification. *arXiv preprint arXiv:2409.06270*.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6202–6211.

Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. 2017. Tall: Temporal activity localization via language query. 5267–5275.

Gao, J.; and Xu, C. 2021. Fast video moment retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1523–1532.

Holmquist, K.; Klasén, L.; and Felsberg, M. 2023. Evidential deep learning for class-incremental semantic segmentation. In *Scandinavian Conference on Image Analysis*, 32–48. Springer.

Honnibal, M.; and Montani, I. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Hu, G.; Xin, Y.; Lyu, W.; Huang, H.; Sun, C.; Zhu, Z.; Gui, L.; Cai, R.; Cambria, E.; and Seifi, H. 2024. Recent trends of multimodal affective computing: A survey from NLP perspective. *arXiv preprint arXiv:2409.07388*.

Hu, Y.; Liu, M.; Su, X.; Gao, Z.; and Nie, L. 2021. Video moment localization via deep cross-modal hashing. *IEEE Transactions on Image Processing*, 30: 4667–4677.

Huang, H.; Qiao, X.; Chen, Z.; Chen, H.; Li, B.; Sun, Z.; Chen, M.; and Li, X. 2024a. CREST: Cross-modal Resonance through Evidential Deep Learning for Enhanced Zero-Shot Learning. arXiv:2404.09640.

Huang, H.; Qin, C.; Liu, Z.; Ma, K.; Chen, J.; Fang, H.; Ban, C.; Sun, H.; and He, Z. 2024b. Trusted unified feature-neighborhood dynamics for multi-view classification. *arXiv preprint arXiv:2409.00755*.

Huang, H.; Shi, J.; Liu, Z.; Chen, H. H.; Fang, H.; Sun, H.; and He, Z. 2025. Structure-Aware Prototype Guided Trusted Multi-View Classification. *arXiv preprint arXiv:2511.21021*.

Huang, J.; Jin, H.; Gong, S.; and Liu, Y. 2022. Video activity localisation with uncertainties in temporal boundary. In *European Conference on Computer Vision*, 724–740. Springer.

Jøsang, A. 2016. *Subjective logic*, volume 3. Springer.

Lan, X.; Yuan, Y.; Chen, H.; Wang, X.; Jie, Z.; Ma, L.; Wang, Z.; and Zhu, W. 2023. Curriculum Multi-Negative Augmentation for Debiased Video Grounding. In *AAAI Conference on Artificial Intelligence*.

Lan, X.; Yuan, Y.; Wang, X.; Chen, L.; Wang, Z.; Ma, L.; and Zhu, W. 2022. A Closer Look at Debiased Temporal Sentence Grounding in Videos: Dataset, Metric, and Approach. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19: 1 – 23.

Lei, J.; Berg, T. L.; and Bansal, M. 2021a. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34: 11846–11858.

Lei, J.; Berg, T. L.; and Bansal, M. 2021b. Detecting Moments and Highlights in Videos via Natural Language Queries. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 11846–11858. Curran Associates, Inc.

Lei, J.; Berg, T. L.; and Bansal, M. 2021c. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34: 11846–11858.

Li, P.; Xie, C.-W.; Xie, H.; Zhao, L.; Zhang, L.; Zheng, Y.; Zhao, D.; and Zhang, Y. 2024. Momentdiff: Generative video moment retrieval from random to real. *Advances in neural information processing systems*, 36.

Lin, K. Q.; Zhang, P.; Chen, J.; Pramanick, S.; Gao, D.; Wang, A. J.; Yan, R.; and Shou, M. Z. 2023. Univtg: Towards unified video-language temporal grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2794–2804.

Liu, B.; Yeung, S.; Chou, E.; Huang, D.-A.; Fei-Fei, L.; and Niebles, J. C. 2018. Temporal modular networks for retrieving complex compositional activities in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 552–568.

Liu, Y.; He, J.; Li, W.; Kim, J.; Wei, D.; Pfister, H.; and Chen, C. W. 2024a. R2-Tuning: Efficient Image-to-Video Transfer Learning for Video Temporal Grounding. *arXiv preprint arXiv:2404.00801*.

Liu, Y.; Li, S.; Wu, Y.; Chen, C.-W.; Shan, Y.; and Qie, X. 2022. UMT: Unified Multi-Modal Transformers for Joint Video Moment Retrieval and Highlight Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3042–3051.

Liu, Z.; Huang, H.; Letchmunan, S.; and Deveci, M. 2024b. Adaptive weighted multi-view evidential clustering with feature preference. *Knowledge-Based Systems*, 294: 111770.

Liu, Z.; Li, J.; Xie, H.; Li, P.; Ge, J.; Liu, S.-A.; and Jin, G. 2024c. Towards balanced alignment: Modal-enhanced semantic modeling for video moment retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 3855–3863.

Lu, W.; Li, J.; Yu, A.; Chang, M.-C.; Ji, S.; and Xia, M. 2024. Llava-mr: Large language-and-vision assistant for video moment retrieval. *arXiv preprint arXiv:2411.14505*.

Ma, K.; Huang, H.; Chen, J.; Chen, H.; Ji, P.; Zang, X.; Fang, H.; Ban, C.; Sun, H.; Chen, M.; et al. 2024. Beyond uncertainty: Evidential deep learning for robust video temporal grounding. *arXiv preprint arXiv:2408.16272*.

Meinardus, B.; Batra, A.; Rohrbach, A.; and Rohrbach, M. 2024. The surprising effectiveness of multimodal large language models for video moment retrieval. *arXiv e-prints*, arXiv–2406.

Moon, W.; Hyun, S.; Lee, S.; and Heo, J.-P. 2023a. Correlation-guided query-dependency calibration for video temporal grounding. *arXiv preprint arXiv:2311.08835*.

Moon, W.; Hyun, S.; Park, S.; Park, D.; and Heo, J.-P. 2023b. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23023–23033.

Moon, W.; Hyun, S.; Park, S.; Park, D.; and Heo, J.-P. 2023c. Query-Dependent Video Representation for Moment Retrieval and Highlight Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 23023–23033.

Otani, M.; Nakashima, Y.; Rahtu, E.; and Heikkilä, J. 2020a. Uncovering hidden challenges in query-based video moment retrieval. *arXiv preprint arXiv:2009.00325*.

Otani, M.; Nakashima, Y.; Rahtu, E.; and Heikkilä, J. 2020b. Uncovering hidden challenges in query-based video moment retrieval. *arXiv preprint arXiv:2009.00325*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021a. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021b. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Regneri, M.; Rohrbach, M.; Wetzel, D.; Thater, S.; Schiele, B.; and Pinkal, M. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1: 25–36.

Sensoy, M.; Kaplan, L.; and Kandemir, M. 2018. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31.

Shafer, G. 1992. Dempster-shafer theory. *Encyclopedia of artificial intelligence*, 1: 330–331.

Wang, C.; Wang, X.; Zhang, J.; Zhang, L.; Bai, X.; Ning, X.; Zhou, J.; and Hancock, E. 2022a. Uncertainty estimation for stereo matching based on evidential deep learning. *Pattern Recognition*, 124: 108498.

Wang, J.; Ma, L.; and Jiang, W. 2020. Temporally grounding language queries in videos by contextual boundary-aware prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12168–12175.

Wang, Z.; Wang, L.; Wu, T.; Li, T.; and Wu, G. 2022b. Negative Sample Matters: A Renaissance of Metric Learning for Temporal Grounding. In *AAAI*, 2613–2623. AAAI Press.

Wu, J.; Li, G.; Liu, S.; and Lin, L. 2020. Tree-Structured Policy based Progressive Reinforcement Learning for Temporally Language Grounding in Video. *ArXiv*, abs/2001.06680.

Wu, W.; Zhang, C.; and Woodland, P. C. 2023. Estimating the uncertainty in emotion attributes using deep evidential regression. *arXiv preprint arXiv:2306.06760*.

Wu, Y.; Shi, B.; Dong, B.; Zheng, Q.; and Wei, H. 2024. The Evidence Contraction Issue in Deep Evidential Regression: Discussion and Solution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 21726–21734.

Ye, K.; Chen, T.; Wei, H.; and Zhan, L. 2024. Uncertainty Regularized Evidential Regression. *arXiv preprint arXiv:2401.01484*.

Zeng, R.; Xu, H.; Huang, W.; Chen, P.; Tan, M.; and Gan, C. 2020. Dense regression network for video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10287–10296.

Zhang, D.; Dai, X.; Wang, X.; Wang, Y.-F.; and Davis, L. S. 2019a. MAN: Moment Alignment Network for Natural Language Moment Retrieval via Iterative Graph Adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, H.; Sun, A.; Jing, W.; and Zhou, J. T. 2023. Temporal sentence grounding in videos: A survey and future directions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, S.; Peng, H.; Fu, J.; and Luo, J. 2019b. Learning 2D Temporal Adjacent Networks for Moment Localization with Natural Language. In *AAAI Conference on Artificial Intelligence*.

Zhang, S.; Peng, H.; Fu, J.; and Luo, J. 2020. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12870–12877.

Zhang, S.; Su, J.; and Luo, J. 2019. Exploiting temporal relationships in video moment localization with natural language. In *Proceedings of the 27th ACM International Conference on Multimedia*, 1230–1238.

Zhang, Z.; Lin, Z.; Zhao, Z.; and Xiao, Z. 2019c. Cross-modal interaction networks for query-based moment retrieval in videos. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 655–664.