De-biased Natural Language Egocentric Task Verification via Prototypical Evidence Learning

Chong Liu[1], Xun Jiang[1], Fumin Shen[1]*, Lei Zhu[2], Jingkuan Song[2], Heng Tao Shen[2], Xing Xu[2]

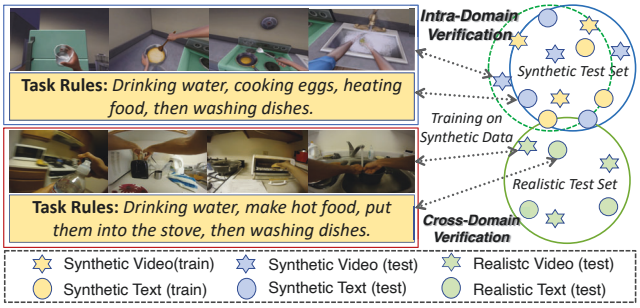
[1]Center for Future Media & School of Computer Science and Engineering,
University of Electronic Science and Technology of China, China
[2]School of Computer Science and Technology, Tongji University, China

Abstract

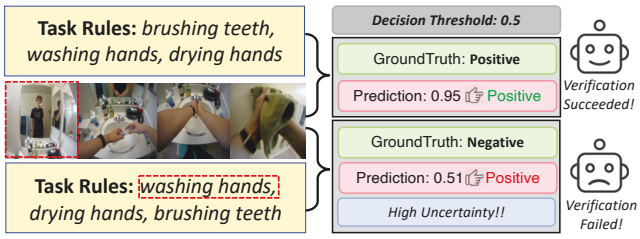
Natural Language-based Egocentric Task Verification (NLETV) aims to verify the alignment between action sequences in egocentric videos and their corresponding textual descriptions. However, existing NLETV approaches are still facing two critical challenges: (1) These methods are designed for simulating environments, ignoring the domain gap between synthetic and realistic data. (2) The matching processes are regarded as a simple binary classification problem, which undermines model reliability due to evaluation bias and uncalibrated decision settings. To address these challenges, we propose a novel method termed **P**rototypical **E**vidential **L**earning (**PEL**), which can be adapted to existing NLETV approaches and boost the model generalization and mitigate prediction bias. Our method leverages prototypes to guide cross-domain alignment and evidence collection. Specifically, PEL consists of two key components: (1) Prototypical Domain Adaptation module enabling cross-domain feature alignment and intra-domain prototype preservation between synthetic and realistic domains; (2) Matching Evidence Collector module, which quantifies prediction uncertainty from the prototypical representations through evidential deep learning. It enforces the model to collect the vision-text consistency and discrepancy evidence, thus addressing the issues of biased decisions in binary classification. Extensive experiments on two public datasets demonstrate that our PEL method outperforms existing state-of-the-art NLETV methods and shows remarkable generalizability.

Introduction

The advancement of robotic systems and wearable technologies has driven a growing demand for egocentric activity recognition (Sener et al. 2022; He et al. 2024; Hutchinson and Gadepally 2021) in untrimmed videos. Within this context, Natural Language-based Egocentric Task Verification (NLETV) (Hazra et al. 2023; Jiang et al. 2025) has emerged as a crucial capability for intelligent systems, enabling them to understand and verify whether executed actions match given natural language instructions. Compared to conventional video-based task verification approaches (Sener et al. 2022; Qian et al. 2022; Dong et al. 2023), NLETV offers superior practicality and generalizability by leveraging natural language descriptions. This paradigm not only aligns with human cognitive patterns but also provides enhanced flexibility in describing diverse procedural tasks.



(a) Domain shift in cross-domain scenario.



(b) Uncalibrated decision setting.

Figure 1: (a) Domain shift in cross-domain scenario. (b) Uncalibrated decision setting.

However, as depicted in Fig. 1, the NLETV task is also highly challenging due to the following two reasons: *(1) Domain shift in cross-domain scenarios.* Existing NLETV methods (Hazra et al. 2023; Jiang et al. 2025) demonstrate strong alignment capabilities for synthetic video-text pairs, benefiting from their training on synthetic data. However, the model performance in real-world scenarios remains suboptimal due to the domain shift, even when processing identical textual descriptions. This performance gap highlights the necessity for cross-domain knowledge learning to enhance model generalization ability in NLETV applications. *(2) Uncalibrated decision setting.* The NLETV task is typically formulated as a binary classification prob-

*Corresponding author.

lem, where the model predicts the alignment probability between a video and its textual description. While the model exhibits sensitivity to textual variations, *e.g.*, altered action sequences, by producing an attenuated probability score, it nevertheless fails to reliably detect the video-text discrepancy. This observation underscores the need for models with calibrated uncertainty estimates to support robust decision-making under ambiguity. To address the aforementioned challenges, we propose **P**rototypical **E**vidential **L**earning (**PEL**), a novel framework that enhances cross-domain generalization while mitigating prediction bias in NLETV tasks.

As illustrated in Fig. 2, PEL consists of two main components: (1) The Prototypical Domain Adaptation (PDA) module. This module employs dual mechanisms to alleviate domain shift: Cross-Domain Alignment, which minimizes distribution discrepancies between source (synthetic) and target (realistic) domains through prototype-guided gradual alignment, preserving both shared cross-domain features and domain-specific characteristics; Intra-Domain Preservation, which computes intra-domain prototype-anchor similarities to maintain task-specific knowledge during adaptation. By leveraging these two mechanisms, PDA ensures that the task semantic structure of each domain is preserved while aligning the features across domains. (2) The Matching Evidence Collector (MEC) module, which leverages evidential deep learning to quantify prediction uncertainty by collecting evidence from prototypes. By explicitly modeling uncertainty, MEC enables the model to distinguish between confident and uncertain cases, facilitating calibrated decision-making and reducing prediction bias. Furthermore, MEC enforces models to collect evidence of both vision-text consistency and discrepancy, tackling the limitations of conventional binary classification settings. We evaluate our proposed PEL method on two benchmark datasets, EgoTV (Hazra et al. 2023) and EgoCross (Jiang et al. 2025). Extensive experiments demonstrate that PEL achieves state-of-the-art performance in NLETV tasks.

In summary, our principal contributions are threefold:

- We propose a novel Prototypical Evidential Learning (PEL) method that addresses the synthetic-to-realistic domain shift and reduces prediction bias in NLETV tasks. Our PEL method can be applied to existing NLETV methods, significantly boosting model performance and generalizability.
- We design a Prototypical Domain Adaptation module, which leverages prototypical video-text joint representations to address the limitations of the domain shift problem in current NLETV models.
- We develop a Matching Evidence Collector module, which quantifies model uncertainty and collects consistency and discrepancy evidence between vision and text, effectively improving decision-making reliability.

Related Work

Natural Language-based Egocentric Task Verification. Egocentric video understanding has evolved from analyzing simple visual patterns to complex cross-modal reasoning, establishing itself as a pivotal research direction. Early foundational works primarily addressed atomic tasks, where models identified predefined activity classes from visual streams. Subsequent advancements extends to cross-modal reasoning (Wang, Huang, and Yuan 2025; Xu et al. 2024; Wang et al. 2024; Jiang et al. 2024b) and multimodal learning (Gao et al. 2024; Hu et al. 2025; Wang et al. 2025), which demanded joint understanding of visual and auxiliary modalities. Recently, Hazra *et al.*introduced the NLETV task (Hazra et al. 2023), which is the focus of our work. The NLETV task demands fine-grained alignment verification between procedural action sequences in videos and their step-by-step textual instructions, representing a more challenging and practical scenario. Furthermore, the release of the EgoCross dataset (Jiang et al. 2025) has enabled cross-domain research in NLETV by providing both synthetic and real-world data, thus facilitating the study of domain adaptation and generalization in this context and inspiring the development of our approach.

Evidential Deep Learning (EDL). The EDL (Sensoy, Kaplan, and Kandemir 2018) techniques have emerged as a promising framework for uncertainty quantification in deep neural networks (Bao, Yu, and Kong 2021; Soleimany et al. 2021; Ulmer, Hardmeier, and Frellsen 2021). Departing from traditional point estimation paradigms, EDL models predictions as evidence that can be accumulated to form Dirichlet-distributed belief masses. This formulation enables models to naturally express uncertainty when evidence is insufficient or conflicting. Recent works have successfully adapted EDL to diverse vision tasks, including Object Detection (Park et al. 2023; Nallapareddy et al. 2023), Action Recognition (Bao, Yu, and Kong 2021; Guo, Wang, and Ji 2024; Zhao et al. 2023) and Temporal Action Localization (Jiang et al. 2022; Gao, Chen, and Xu 2023; Chen et al. 2022; Jiang et al. 2024c,a). These applications consistently demonstrate EDL's capability to improve model reliability in ambiguous cases. Building on these foundations, we propose our PEL method that quantifies prediction uncertainty and collects the consistency and discrepancy evidence of video-text pairs in NLETV tasks.

Method

Preliminaries

Problem Definition. Let an egocentric video be represented as a sequence of frames or short video clips, denoted as $V = \{v_i\}_{i=1}^{n}$, where v_i corresponds to the i-th video segment and n is the total number of segments in the video. The associated natural language description $T = \{t_j\}_{j=1}^{m}$ comprises a sequence of action instructions, with each t_j describing the j-th action step in the procedural task. The NLETV problem is formally defined as learning a parameterized binary classifier $f_\theta : \mathcal{V} \times \mathcal{T} \rightarrow \{0, 1\}$, where $\mathcal{V}$ and $\mathcal{T}$ represent the video and text domains, respectively, and θ denotes the learnable parameters of the model. The classifier's objective is to determine whether the video execution V faithfully aligns with the textual description T. For cross-domain NLETV evaluation, we follow the EgoCross benchmark dataset (Jiang et al. 2025), utilizing fully labeled source domain data alongside limited unlabeled target do-

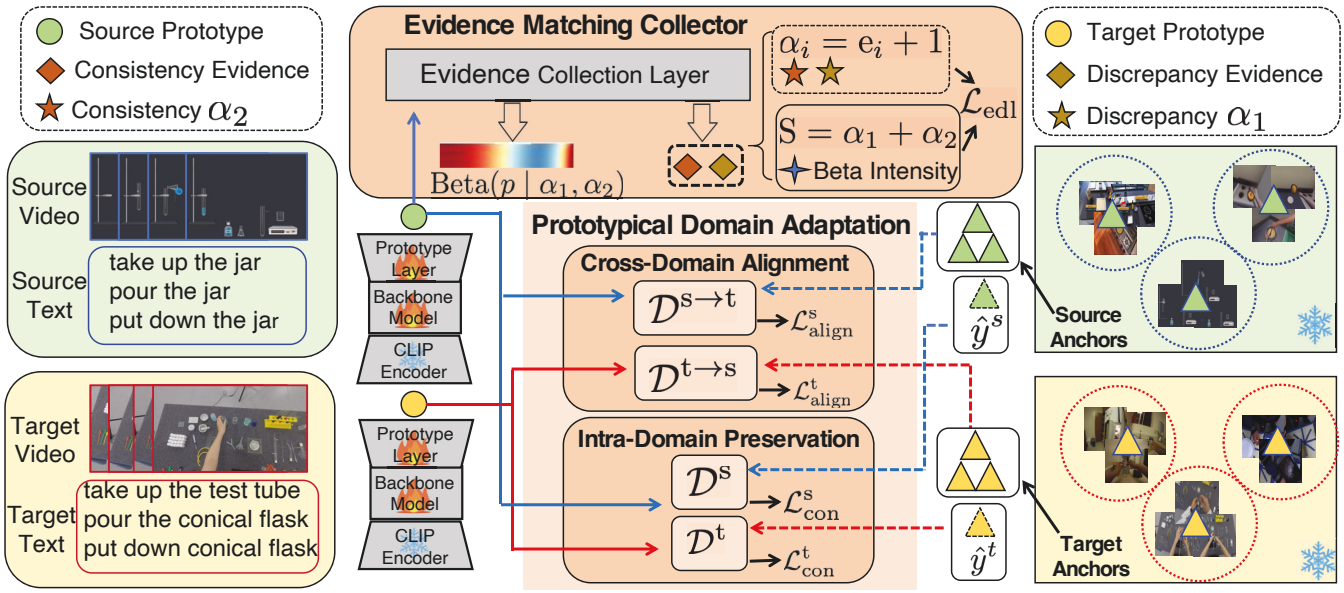


Figure 2: An overview of our PEL method, which consists of two main components: (1) Prototypical Domain Adaptation (PDA) and (2) Matching Evidence Collector (MEC). Note that $\mathcal{D}^{s\to t}$ means alignment from source to target domain, while $\mathcal{D}^{t\to s}$ means alignment from target to source domain.

main samples during training, with final performance assessment conducted exclusively on the target domain test set.

Subject Logic. Evidential Deep Learning (EDL) (Sensoy, Kaplan, and Kandemir 2018) builds upon the theoretical framework of Subjective Logic (Jsang 2018), which provides a principled approach for uncertainty quantification in neural networks. The key innovation lies in its explicit decomposition of model outputs into three interpretable components: belief mass b_k, probability values p_k, and uncertainty u. For NLETV's binary classification setting, given a video-text pair (V, T) processed by model f_θ, we first extract the evidence vectors $\mathbf{e} = [e_1, e_2]$. These evidence values parameterize a Beta distribution through the following transformation:

$$b_k = \frac{e_k}{S}, \quad p_k = \frac{e_k + 1}{S}, \quad u = \frac{2}{S}, \tag{1}$$

where $S = \sum_{k=1}^{2}(e_k + 1)$ represents the Dirichlet intensity (reduced to Beta for binary cases) and u is inversely proportional to the total evidence. The resulting Beta distribution can be expressed as:

$$\text{Beta}(p \mid \alpha_1, \alpha_2) = \frac{1}{B(\alpha_1, \alpha_2)} p^{\alpha_1 - 1}(1-p)^{\alpha_2 - 1}, \tag{2}$$

where $\alpha_k = e_k + 1$, provides a full probability distribution over the alignment probability $p \in [0, 1]$, where p represents the probability of the video-text alignment.

Prototypical Domain Adaptation

We propose the Prototypical Domain Adaptation (PDA) module, which not only leverages prototypes to guide cross-domain feature alignment but also preserves the inherent task structure of each domain. PDA module consists of two key components: (1) Cross-Domain Alignment (CDA) and (2) Intra-Domain Preservation (IDP).

Task Anchors Generation. We utilize CLIP (Radford et al. 2021) to extract semantically meaningful features, taking advantage of its cross-modal representation space trained on large-scale video-text data. This pretraining endows the model with a strong ability to capture cross-modal semantic relationships. For video feature processing, we first aggregate frame-level features $F = \{f_i\}_{i=1}^{n}$ into a compact video-level representation $F' \in \mathbb{R}^d$ via pooling operations. We then perform K-Means clustering (McQueen 1967) on the aggregated features across the training set to discover K semantic anchors:

$$\mathcal{A} = \{\mathbf{a}_k\}_{k=1}^{K} = \text{K-means}(\{\mathbf{F'}_j\}_{j=1}^{M}), \tag{3}$$

where M denotes the total training videos. For frame-level assignment, we compute the similarity between each frame-level feature f_i and anchor set $\mathcal{A}$ to obtain frame-cluster assignments c_i. The final video-level anchor assignment y_j is determined through majority voting:

$$y_j = \underset{k \in \{1,\ldots,K\}}{\arg\max} \sum_{i=1}^{n} \mathbb{I}(c_{ji} = k). \tag{4}$$

These anchor assignments y_j serve as pseudo-labels that capture the underlying task semantic structure of the video, guiding both cross-domain alignment and intra-domain prototype preservation in subsequent training stages.

Prototypes Generation. We construct domain-specific prototypes by aggregating multimodal features from video-text pairs in both source and target domains. Specifically, Video frames and text descriptions are encoded using CLIP's pretrained vision and language encoders to obtain frame-level

features $\mathbf{v}_i \in \mathbb{R}^{d_v}$ and sentence-level features $\mathbf{t}_j \in \mathbb{R}^{d_t}$, respectively. These unimodal features are then fused by the NLETV backbone to produce a joint representations $\mathbf{f} \in \mathbb{R}^{d_f}$ that capture cross-modal interactions. To obtain discriminative prototypes, we design a prototype layer that performs linear projection, which learns a prototype set $\{\mathbf{p}_i\}_{i=1}^{N}$ to serve as compact aggregations of video-text semantics and are subsequently used to reduce domain shift and to effectively support evidence aggregation.

Cross-Domain Alignment. Our Cross-Domain Alignment (CDA) mechanism effectively facilitates cross-domain knowledge transfer by establishing probabilistic correspondences between domain-specific prototypes and task anchors from the opposite domain. This mechanism enables the model to capture domain-invariant semantics while respecting domain-specific nuances. The alignment process begins by generating domain-specific prototypes through video-text feature aggregation. For each prototype $\mathbf{p}_i$ in one domain, we compute its alignment probabilities with each of the K task anchors $\{\mathbf{a}_k\}_{k=1}^{K}$ from the opposite domain using a temperature-scaled softmax over cosine similarities:

$$p_{ik} = \frac{\exp\left(\tau \cdot \cos(\mathbf{p}_i, \mathbf{a}_k)\right)}{\sum_{j=1}^{K} \exp\left(\tau \cdot \cos(\mathbf{p}_i, \mathbf{a}_j)\right)}, \tag{5}$$

where τ is a temperature scaling factor that controls the sharpness of the similarity distribution. To enforce confident and meaningful cross-domain alignment, we employ a bidirectional entropy minimization strategy. Specifically, for source-to-target alignment, we minimize the entropy of the alignment probabilities:

$$\mathcal{L}_{\text{align}}^{\text{s}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} p_{ik}^{\text{s}\to\text{t}} \log(p_{ik}^{\text{s}\to\text{t}} + \epsilon), \tag{6}$$

and symmetrically for target-to-source alignment:

$$\mathcal{L}_{\text{align}}^{\text{t}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} p_{ik}^{\text{t}\to\text{s}} \log(p_{ik}^{\text{t}\to\text{s}} + \epsilon), \tag{7}$$

where ϵ is a small constant to avoid numerical instability. The composite alignment loss $\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{align}}^{\text{s}} + \mathcal{L}_{\text{align}}^{\text{t}}$ encourages the model to capture the shared knowledge across domains while preserving domain-specific characteristics.

Intra-Domain Preservation. To maintain domain-specific task structures during adaptation, we introduce an intra-domain preservation mechanism that encourages consistency between prototypes and their corresponding task anchors within each domain. Given the domain-specific prototypes $\mathbf{p}_i$ and the task anchors $\{\mathbf{a}_k\}_{k=1}^{K}$, the mechanism operates in two stages. First, we compute cosine similarities between each prototype and all task anchors in the same domain, and assign each prototype to the anchor with the highest similarity:

$$\hat{y} = \arg\max_{k} \cos(\mathbf{p}_i, \mathbf{a}_k). \tag{8}$$

These predicted anchor assignments, denoted as $\hat{y}^s$ for the source domain and $\hat{y}^t$ for the target domain, are then compared to the ground-truth cluster labels y^s and y^t (obtained from Eq. 4) using a cross-entropy loss:

$$\mathcal{L}_{\text{con}} = \text{CE}(\hat{y}_{\text{s}}, y_{\text{s}}) + \text{CE}(\hat{y}_{\text{t}}, y_{\text{t}}). \tag{9}$$

This regularization ensures that the alignment of prototypes to task anchors remains faithful to the original domain semantics, thereby preserving task-specific knowledge and structural integrity during domain adaptation.

Matching Evidence Collector

We propose the Matching Evidence Collector (MEC) module, which quantifies prototypical prediction uncertainty while explicitly collecting both consistent and discrepant cross-modal evidence to enhance decision reliability. Specifically, for processing prototype features f, we design an evidence collection layer to map high-dimensional features to binary classification space f', which is used to collect both consistent and discrepant evidence. Using exponential transformation, we derive evidence vectors $e = [e_1, e_2]$ where $e_k = \exp(f'_k)$, with e_1 representing consistency evidence and e_2 representing discrepancy evidence. To regulate evidence accumulation and uncertainty estimation, we define two key components of the Dirichlet distribution: total intensity $S = \sum_{i=1}^{2}(e_i + 1)$ and uncertainty $u = \frac{2}{S}$. We first leverage the classification loss to encourage alignment between predicted evidence and ground-truth labels while regularizing prediction variance:

$$\mathcal{L}_{\text{cls}} = \sum_{k=1}^{2} \left[(y_k - \frac{\alpha_k}{S})^2 + \frac{\alpha_k(S - \alpha_k)}{S^2(S+1)} \right], \tag{10}$$

where $\alpha_k = e_k + 1$ denotes the Dirichlet parameter for class k. To mitigate overconfidence of low-quality data(*e.g.*blurry videos), we introduce a KL term to regularizes the evidence:

$$\mathcal{L}_{\text{kl}} = \sum_{k=1}^{2} \left[(\alpha_k - 1)(\psi(\alpha_k) - \psi(S)) + \ln \frac{\Gamma(S)}{\Gamma(\alpha_k)} \right], \tag{11}$$

where $\psi(\cdot)$ and $\Gamma(\cdot)$ denote the digamma and gamma functions, respectively. The overall evidential loss is then defined as: $\mathcal{L}_{\text{edl}} = \mathcal{L}_{\text{cls}} + \lambda \mathcal{L}_{\text{kl}}$. where λ is a hyperparameter that balances the contributions of the KL regularization term. Minimizing this loss encourages MEC to quantify uncertainty from prototypical representations and collect both consistent and discrepant evidence, thereby resolving the bias issues inherent to binary classification in NLETV tasks.

Overall Training

In a nutshell, we optimize the framework by minimizing a combined loss function that integrates the key loss components introduced earlier, which can be formulated as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{edl}} + \lambda_1 \mathcal{L}_{\text{align}} + \lambda_2 \mathcal{L}_{\text{con}}, \tag{12}$$

where $\mathcal{L}_{\text{edl}}$ regulates uncertainty quantification and evidence collection; $\mathcal{L}_{\text{align}}$ enforces to capture the shared knowledge across domains and preserve domain-specific characteristics; and $\mathcal{L}_{\text{con}}$ maintains intra-domain consistency of task-specific knowledge; λ_1 and λ_2 are hyperparameters that balance the contributions of each loss term.

Method	Lab			Kitchen			Daily			Average
	Novel Tasks	Novel Steps	Omitted Steps	Novel Tasks	Novel Steps	Omitted Steps	Novel Tasks	Novel Steps	Omitted Steps	
GoldFish (ECCV'24)	43.4	41.8	44.3	53.1	48.3	45.6	58.9	48.1	50.9	48.3
ShareGPT4Video (NeurIPS'24)	47.9	40.1	40.5	52.9	47.2	46.8	60.6	57.0	50.1	49.2
GroundingGPT (ACL'24)	34.0	35.6	34.3	40.3	37.6	38.0	34.3	35.1	34.8	36.0
Otter (TPAMI'25)	44.6	43.2	46.5	52.3	47.8	44.3	57.6	49.3	49.9	48.4
NSG (ICCV'23)	32.4	34.3	35.8	39.7	72.4	40.1	84.0	82.5	43.4	51.7
NSG+PEL (Ours)	**48.8**	**64.8**	**44.6**	**57.5**	**79.2**	**44.7**	**84.9**	**87.1**	**46.2**	**61.9**(↑10.2)
PHGC (CVPR'25)	33.2	37.3	34.9	35.7	79.2	44.1	84.1	84.4	47.3	53.4
PHGC+PEL (Ours)	**52.8**	**62.4**	**47.1**	**73.3**	**84.2**	**47.4**	**85.1**	**86.1**	**57.2**	**66.2**(↑12.8)

Table 1: Performance comparison on the EgoCross dataset in terms of Macro-F1. MLLMs are highlighted in gray, and the best results are shown in **bold**.

Experiments

Implementation Details

Datasets and Metrics. We implemented our method on two datasets: *(1) EgoTV* (Hazra et al. 2023): it contains task descriptions and video pairs labeled as match or mismatch, collected from digital environments. It includes 130 target objects and 24 receptacle objects, resulting in 1038 unique task-object combinations performed across 30 different kitchen scenes. The dataset covers 82 tasks, and is split into 10726, 1080, 700, 2164 and 676 for training, novel tasks, novel steps, novel scenes and abstraction respectively. *(2) EgoCross* (Jiang et al. 2025): The public EgoCross dataset contains both synthetic and real video-text pairs and is composed of three sub-datasets: Lab Experiments, Kitchen Cooking, and Daily Life. It includes a total of 13,990 video-text pairs covering 156 distinct tasks. Approximately 70% of the data is used for training, and the remaining 30% is divided into three separate test sets: Novel Tasks, Novel Steps, and Omitted Steps.

Experimental Settings. Following prior work (Jiang et al. 2025), we utilize pre-trained models as feature extractors to obtain visual and textual representations offline for the EgoTV dataset. Specifically, CLIP (Radford et al. 2021) servers as the backbone for extracting frame-level features from videos and encoding textual descriptions into high-dimensional embeddings. For the video modality, 20 frames are sampled from each video. For the text modality, we first adopt the pre-trained T5 model (Raffel et al. 2020), as in (Hazra et al. 2023), to extract object states and operational topological vertices.

De-biased Evaluation Metrics. The NLETV task demands that the model accurately verify the matching relationship between videos and textual descriptions, where both match and mismatch classes hold equal significance. However, prior works commonly adopt the F1-score (Powers 2020) as the evaluation metric, which inherently favors the match class and introduces class bias (as reflected by Fig. 3). To analyze this issue, we evaluate five methods, *i.e.*, Random Prediction (Random), All Match (Default-T), NSG and PHGC, on the EgoTV dataset. The results shows that F1-score of the match and mismatch classes are imbalanced, with higher F1-score for the match class in both NSG and PHGC methods. This bias highlights the inadequacy of F1-score in providing a fair evaluation. To address this limitation, we adopt the Macro-F1 metric (Opitz and Burst 2019), which treats both classes on an equal footing and effectively alleviate the bias present in previous evaluation protocols. Notably, all experiments conducted on the EgoTV and EgoCross datasets utilize this metric to ensure fair and unbiased performance assessment.

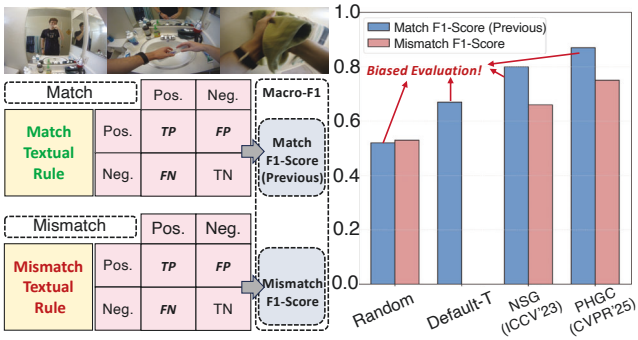


Figure 3: An illustration of evaluation bias due to F1-Score in previous works.

Overall Comparison

We apply our PEL method to the NSG and PHGC baselines (denoted as NSG+PEL and PHGC+PEL) for evaluation on the EgoCross dataset. For the EgoTV dataset, we incorporate our MEC module into NSG and PHGC to form enhanced variants. Following the evaluation protocol in (Jiang et al. 2025), we compare our method with state-of-the-art NLETV methods: PHGC (Jiang et al. 2025) and NSG (Hazra et al. 2023). Furthermore, we also evaluate four Multimodal Large Language Models (MLLMs) applicable to NLETV tasks: GoldFish (Ataallah et al. 2024), Otter (Li et al. 2025), ShareGPT4Video (Chen et al. 2024), and GroundingGPT (Li et al. 2024).

For clarity, we use the following abbreviations to denote evaluation scenarios: NT (Novel Tasks), NS (Novel Steps), NSC (Novel Scenes), ABS (Abstraction), Omitted Steps (OS), and AVG (Average). Experimental results are presented in Table 1 and Table 2. By comparing the performance of our approach with other models, we summarize the following observations: (1) Our PEL method significantly

Method	NT	NS	NSC	ABS	AVG
GoldFish (ECCV'24)	44.8	54.4	44.4	52.6	49.1
Otter (TPAMI'25)	50.4	60.1	56.3	63.3	57.5
ShareGPT4Video (NeurIPS'24)	49.3	51.6	52.1	47.2	50.1
GroundingGPT (ACL'24)	35.4	36.7	35.6	37.8	36.4
NSG (ICCV'23)	80.3	65.7	76.6	72.9	73.9
NSG+MEC (Ours)	**82.1**	**68.5**	**80.2**	**76.4**	**76.8**
PHGC (CVPR'25)	88.4	74.3	80.0	82.2	81.2
PHGC+MEC (Ours)	**90.3**	**77.4**	**83.8**	**83.7**	**83.8**

Table 2: Performance comparison on the EgoTV dataset in terms of Macro-F1. MLLMs are highlighted in gray.

outperforms all baselines on the EgoCross dataset. Specifically, NSG+PEL and PHGC+PEL achieve average Macro-F1 scores of 61.9% and 66.2%, representing absolute improvements of 10.2% and 12.8% compared to their original counterparts, respectively. This consistent enhancement validates the effectiveness of PEL in mitigating both the domain gap between synthetic and realistic data and the prediction bias inherent in NLETV tasks. (2) On the synthetic EgoTV dataset, integrating MEC into NSG and PHGC yields average Macro-F1 scores of 76.8% and 83.8%, with incremental improvements of 2.9% and 2.6%. These results demonstrate that MEC retains its capability to alleviate classification bias even in structured synthetic data scenarios, where the domain shift is inherently minimal.

However, we acknowledge certain limitations of our method. On the EgoCross dataset, our method does not exhibit a significant advantage over MLLMs in the Omitted Steps metric. This subtask involves verifying video-text pairs with one or more missing actions, demanding fine-grained procedural reasoning capabilities. We hypothesize that this is because our approach primarily focuses on mitigating domain shift and bias, whereas MLLMs benefit from their complex model architectures and extensive pre-training on large-scale synthetic and realistic data, resulting in stable performance across all three subtasks. Nevertheless, it is noteworthy that our method still achieves state-of-the-art results on both datasets,confirming its practical value for NLETV research.

Further Study

Analysis on Number of Anchors. Notably, setting K to 40 yields optimal performance, as this configuration balances semantic coverage and discriminative power. At K = 40, anchors effectively span the EgoCross dataset's semantic space, capturing both coarse-grained task categories and fine-grained details—critical for robust cross-domain alignment and intra-domain preservation. When K is too small (*e.g.*, 30), anchors fail to cover the full range of task-specific information,leading to under-represented concepts and imprecise alignment. Conversely, a larger K (*e.g.*, 50) introduces redundant, overlapping anchors in the semantic space, increasing matching complexity and noise that hinder performance. Thus, $K = 40$ strikes an optimal balance, ensuring robust domain adaptation and reliable evidence collection.

Ablation Study. Based to the ablation study results presented in Table 4, we draw the following key conclusions:

Cluster	Lab				Kitchen			
	NT	NS	NSC	AVG	NT	NS	NSC	AVG
K=0	33.2	37.3	34.9	35.1	41.4	77.1	40.6	53.0
K=20	40.3	56.3	42.7	46.4	70.3	83.5	44.4	66.1
K=30	36.9	60.9	46.3	48.0	68.2	80.3	41.9	63.5
K=40	**52.8**	**62.4**	47.1	**54.1**	**73.3**	84.2	**47.4**	**68.3**
K=50	38.7	61.7	**47.7**	49.4	71.6	81.1	45.3	66.0
K=60	50.3	40.6	46.4	45.8	70.1	**84.6**	39.7	64.8

Table 3: Analysis on the number of anchors K on Lab and Kitchen subsets of EgoCross dataset.

(1) The Cross-Domain Alignment (CDA) module is indispensable to the overall model performance. Removing the CDA module resulting in a significant performance decline across all metrics, highlighting its critical role in mitigating domain shift. This finding also validates the rationality and effectiveness of leveraging prototypes to guide the learning of cross-domain shared knowledge in the NLETV task. (2) The Intra-Domain Preservation (IDP) module is also crucial for optimizing model performance. By comparing the second and fourth rows, a substantial improvement is observed in the Novel Tasks metric. This indicate that the IDP module effectively captures task-specific information, thereby enhancing the model's generalization capability across diverse tasks scenarios. (3) Enabling the MEC module further boosts model performance by facilitating both consistency and discrepancy evidence collection, which reduces the decision bias inherent in binary classification settings and provides a more comprehensive basis for matching verification.

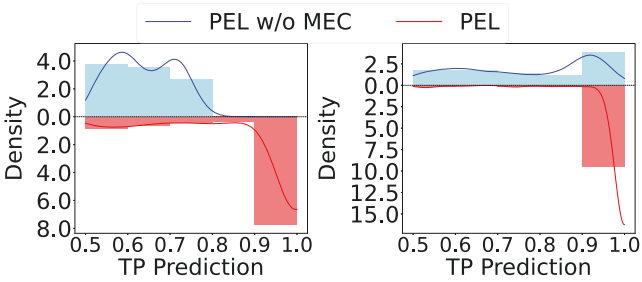


Figure 4: Analysis on TP predictions on Lab and Daily subsets of EgoCross dataset.

Analysis on Matching Evidence Collector. We further evaluate the effectiveness of the Matching Evidence Collector (MEC) module by visualizing the density distribution of true positive (TP) predictions on the Lab and Daily subsets of the EgoCross dataset, as shown in Fig. 4. On the Lab subset, the TP predictions from PHGC+PEL are highly concentrated in the confidence range [0.9,1.0], indicating strong model certainty. In contrast, the model without MEC shows a broader distribution, spanning from 0.5 to 0.8, which suggests uncertainty in its predictions. On the Daily subset, a sample scenario, both models tend to produce high-confidence predictions. However, the model variant without MEC still yields a notable number of low-confidence predictions from 0.5 to 0.8. In comparison, the complete model with MEC almost consistently outputs predictions within the

CDA	IDP	MEC	Lab			Kitchen			Daily			Average
			Novel Tasks	Novel Steps	Omitted Steps	Novel Tasks	Novel Steps	Omitted Steps	Novel Tasks	Novel Steps	Omitted Steps	
-	-	-	33.2	37.3	34.9	35.7	79.2	44.1	80.1	81.4	47.3	52.6
✔	-	-	41.4	55.8	43.6	60.5	82.3	45.5	82.4	85.7	54.6	61.3
-	✔	-	46.1	40.5	40.8	67.3	80.7	44.9	84.3	82.4	50.7	59.7
✔	✔	-	47.4	59.1	46.5	69.6	82.8	45.6	83.8	85.9	54.9	64.0
-	-	✔	36.7	39.5	38.1	41.4	80.1	44.6	82.2	81.9	49.7	54.9
✔	-	✔	43.2	57.6	44.4	64.3	83.4	47.0	83.6	85.8	55.9	62.8
-	✔	✔	49.5	46.6	43.7	69.1	82.7	46.2	84.8	83.3	52.4	62.0
✔	✔	✔	**52.8**	**62.4**	**47.1**	**73.3**	**84.2**	**47.7**	**85.1**	**86.1**	**57.2**	**66.2**

Table 4: Ablation studies regarding key components in our proposed PEL method on the EgoCross benchmark dataset.

[0.9,1.0] range, with the majority clustered near 1.0. These findings suggest that the MEC module plays a crucial role in quantifying prediction uncertainty and aggregating evidence from prototypes, thereby enhancing the robustness and reliability of the model's decisions.

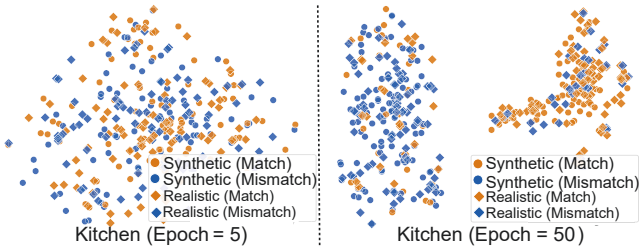


Figure 5: t-SNE visualization of prototypes on Kitchen subsets of EgoCross dataset.

Analysis on Prototype Visualization. To further validate the effectiveness of our proposed PEL module, we perform t-SNE visualizations on the Kitchen subsets at epoch 5 and epoch 50, as presented in Fig. 5. At epoch 5, the data points are scattered and poorly clustered, indicating that the model's predictions remain highly uncertain and fail to effectively distinguish between classes. In contrast, at epoch 50, the source and target domain samples exhibit strong integration, suggesting that the model has learned shared information across domains. Furthermore, the two label clusters are clearly separated, demonstrating high confidence in classification. These visual results collectively confirm that our PEL method effectively leverages prototypes to guide cross-domain knowledge learning and evidence collection, thereby mitigating both domain shift and prediction bias in NLETV tasks.

Qualitative Analysis. We visualize real-world cases from EgoCross to qualitatively evaluate the effectiveness of our PEL method. As illustrated in Fig. 6, our method accurately identifies whether video and text descriptions are aligned, while comparative methods, PHGC and GoldFish, generate erroneous predictions. Specifically, our method correctly predicts the alignment for real-world video-text pairs, whereas PHGC and GoldFish struggle with domain shift between synthetic training data and real-world test scenarios. Furthermore, when the text description remains unchanged

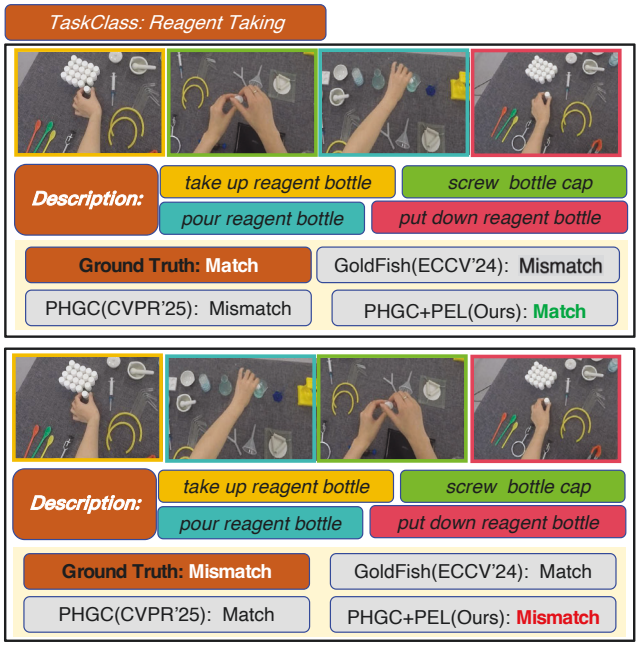


Figure 6: Qualitative analysis of the task verification on the *Lab* scenario of EgoCross dataset.

but the order of video segments is altered, our method again delivers accurate predictions, while the other two methods fail to capture this structural variation.

Conclusion

In this work, we focused on the NLETV task and proposed a novel Prototypical Evidential Learning (PEL) framework. Specifically, we introduced a Prototypical Domain Adaptation module to capture shared knowledge across domains while preserving task-specific information, mitigating domain shift. Additionally, the Matching Evidence Collector module was designed to enhance decision reliability by collecting video-text consistency and discrepancy evidence. Extensive experiments demonstrated that our method outperformed state-of-the-art baselines on both the EgoCross and EgoTV datasets.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants, China (No.62476201, 62222203 and 62306065), the New Cornerstone Science Foundation through the XPLORER PRIZE, and the Central Government Guiding Funds for Local Science and Technology Development, Shanghai, No. YDZX20253100002004.

References

Ataallah, K.; Shen, X.; Abdelrahman, E.; Sleiman, E.; Zhuge, M.; Ding, J.; Zhu, D.; Schmidhuber, J.; and Elhoseiny, M. 2024. Goldfish: Vision-language understanding of arbitrarily long videos. In *European Conference on Computer Vision*, 251–267. Springer.

Bao, W.; Yu, Q.; and Kong, Y. 2021. Evidential deep learning for open set action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 13349–13358.

Chen, L.; Wei, X.; Li, J.; Dong, X.; Zhang, P.; Zang, Y.; Chen, Z.; Duan, H.; Tang, Z.; Yuan, L.; et al. 2024. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37: 19472–19495.

Chen, M.; Gao, J.; Yang, S.; and Xu, C. 2022. Dual-evidential learning for weakly-supervised temporal action localization. In *European conference on computer vision*, 192–208. Springer.

Dong, S.; Hu, H.; Lian, D.; Luo, W.; Qian, Y.; and Gao, S. 2023. Weakly supervised video representation learning with unaligned text for sequential videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2437–2447.

Gao, J.; Chen, M.; and Xu, C. 2023. Collecting cross-modal presence-absence evidence for weakly-supervised audio-visual event perception. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18827–18836.

Gao, Z.; Jiang, X.; Xu, X.; Shen, F.; Li, Y.; and Shen, H. T. 2024. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 26876–26885.

Guo, H.; Wang, H.; and Ji, Q. 2024. Bayesian evidential deep learning for online action detection. In *European Conference on Computer Vision*, 283–301. Springer.

Hazra, R.; Chen, B.; Rai, A.; Kamra, N.; and Desai, R. 2023. Egotv: Egocentric task verification from natural language task descriptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15417–15429.

He, T.; Liu, H.; Li, Y.; Ma, X.; Zhong, C.; Zhang, Y.; and Lin, W. 2024. Collaborative weakly supervised video correlation learning for procedure-aware instructional video analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 2112–2120.

Hu, D.; Jiang, X.; Sun, Z.; Yang, H.; Peng, C.; Yan, P.; Shen, H. T.; and Xu, X. 2025. Geometric Gradient Divergence Modulation for Imbalanced Multimodal Learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 1337–1345.

Hutchinson, M. S.; and Gadepally, V. N. 2021. Video action understanding. *IEEE Access*, 9: 134611–134637.

Jiang, X.; Huang, Z.; Xu, X.; Song, J.; Shen, F.; and Shen, H. T. 2025. PHGC: Procedural Heterogeneous Graph Completion for Natural Language Task Verification in Egocentric Videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 8615–8624.

Jiang, X.; Wei, Z.; Li, S.; Xu, X.; Song, J.; and Shen, H. T. 2024a. Counterfactually augmented event matching for debiased temporal sentence grounding. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 6472–6481.

Jiang, X.; Xu, X.; Zhang, J.; Shen, F.; Cao, Z.; and Shen, H. T. 2022. Sdn: Semantic decoupling network for temporal language grounding. *IEEE Transactions on Neural Networks and Learning Systems*, 35(5): 6598–6612.

Jiang, X.; Xu, X.; Zhou, Z.; Yang, Y.; Shen, F.; and Shen, H. T. 2024b. Zero-shot video moment retrieval with angular reconstructive text embeddings. *IEEE Transactions on Multimedia*.

Jiang, X.; Xu, X.; Zhu, L.; Sun, Z.; Cichocki, A.; and Shen, H. T. 2024c. Resisting noise in pseudo labels: Audible video event parsing with evidential learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Jsang, A. 2018. *Subjective Logic: A formalism for reasoning under uncertainty*. Springer Publishing Company, Incorporated.

Li, B.; Zhang, Y.; Chen, L.; Wang, J.; Pu, F.; Cahyono, J. A.; Yang, J.; Li, C.; and Liu, Z. 2025. Otter: A multi-modal model with in-context instruction tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Li, Z.; Xu, Q.; Zhang, D.; Song, H.; Cai, Y.; Qi, Q.; Zhou, R.; Pan, J.; Li, Z.; Vu, V. T.; et al. 2024. Groundinggpt: Language enhanced multi-modal grounding model. *arXiv preprint arXiv:2401.06071*.

McQueen, J. B. 1967. Some methods of classification and analysis of multivariate observations. In *Proc. of 5th Berkeley Symposium on Math. Stat. and Prob.*, 281–297.

Nallapareddy, M. R.; Sirohi, K.; Drews, P. L.; Burgard, W.; Cheng, C.-H.; and Valada, A. 2023. EvCenterNet: Uncertainty estimation for object detection using evidential learning. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 5699–5706. IEEE.

Opitz, J.; and Burst, S. 2019. Macro f1 and macro f1. *arXiv preprint arXiv:1911.03347*.

Park, Y.; Choi, W.; Kim, S.; Han, D.-J.; and Moon, J. 2023. Active learning for object detection with evidential deep learning and hierarchical uncertainty aggregation. In *The Eleventh International Conference on Learning Representations*.

Powers, D. M. 2020. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

Qian, Y.; Luo, W.; Lian, D.; Tang, X.; Zhao, P.; and Gao, S. 2022. Svip: Sequence verification for procedures in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19890–19902.

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

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

Sener, F.; Chatterjee, D.; Shelepov, D.; He, K.; Singhania, D.; Wang, R.; and Yao, A. 2022. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21096–21106.

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

Soleimany, A. P.; Amini, A.; Goldman, S.; Rus, D.; Bhatia, S. N.; and Coley, C. W. 2021. Evidential deep learning for guided molecular property prediction and discovery. *ACS central science*, 7(8): 1356–1367.

Ulmer, D.; Hardmeier, C.; and Frellsen, J. 2021. Prior and posterior networks: A survey on evidential deep learning methods for uncertainty estimation. *arXiv preprint arXiv:2110.03051*.

Wang, Y.; Huang, W.; and Yuan, C. 2025. Aligning Composed Query with Image via Discriminative Perception from Negative Correspondences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 8078–8086.

Wang, Y.; Liu, L.; Yuan, C.; Li, M.; and Liu, J. 2024. Negative-sensitive framework with semantic enhancement for composed image retrieval. *IEEE Transactions on Multimedia*, 26: 7608–7621.

Wang, Z.; Xu, X.; Zhu, L.; Bin, Y.; Wang, G.; Yang, Y.; and Shen, H. T. 2025. Evidence-Based Multi-Feature Fusion for Adversarial Robustness. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xu, J.; Huang, Y.; Hou, J.; Chen, G.; Zhang, Y.; Feng, R.; and Xie, W. 2024. Retrieval-augmented egocentric video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13525–13536.

Zhao, C.; Du, D.; Hoogs, A.; and Funk, C. 2023. Open set action recognition via multi-label evidential learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22982–22991.