Appendix

Organization of Appendix. In this Appendix, we first summarize main notations. Next, we provide the theoretical proof for Lemma 1. We present additional experimental results and qualitative examples to demonstrate how the tag prediction contributes to the diversity of generated captions. We then provide additional discussion on the choice of datasets, along with the limitations of this work.

Summary of Notations

We summarize the major notations used throughout the paper in Table 4.

Table 4: Summary of Notations

Notation	Description
b_k	Belief mass for class k in evidential learning
vac	Vacuity
V_t	Feature embeddings of video at timestep t
$\mathbf{e}_{i,k}$	Evidence for word k at position i
$\mathbf{a}_{i,k}$	Opinion for word k at position i
S_i	Sum of all opinions at position i
w_i	Word at position i of the caption
$g_{i,k}$	Predicted logit for word k at position i
z	Tag
T	Temperature parameter
K	Size of vocabulary
$y_{n,k}$	Whether word k is the correct word in n-th caption
p_k	Predicted probability of word k to be the next word
c_k	The occurrence of a specific word k from all appropriate captions
A	Set of appropriate words for the next word in caption
N	Number of captions

Proof for Lemma 1

Cross-entropy with multiple words acceptable. Caption generation is usually formulated as a sequential prediction problem where the next word is generated given the previous words. For simplification, we consider the task of predicting the next word as a multi-class classification problem.

Typically, there are multiple correct captions to describe a video. Therefore, given an incomplete sentence fragment, the next word may have multiple appropriate candidates. For a single video, denote the number of ground-truth captions starting with an incomplete sentence fragment as N, and the vocabulary size as K. During training, the commonly used cross-entropy loss is aggregated for all candidate words and all captions:

$$L = -\sum_{n=1}^{N}\sum_{k=1}^{K} y_{nk} \ln p_k \tag{17}$$

We further denote c_k as the occurrence of a specific word k from the N captions, and denote the set A as all appropriate words as

$$\begin{aligned} c_k &= \sum_{n=1}^{N} \delta(y_{nk} = 1) \\ A &= \{k\}, \quad \text{s.t. } c_k > 0 \end{aligned} \tag{18}$$

The the aggregated loss can be rewritten as

$$\begin{aligned} L = -\sum_{k=1}^{V} c_k \ln p_k = -\sum_{k\in A} c_k \ln p_k \\ \text{s.t.} \sum_{k=1}^{V} p_k = 1, \quad \forall k : 0 \le k \le 1 \end{aligned} \tag{19}$$

To minimize the aggregated loss, we first take the exponential as

$$\exp(-L) = \prod (p_k)^{c_k} \tag{20}$$

This form matches the probability mass function of multinomial distribution, and the maximum likelihood estimation is straightforward

$$\hat{p}_k = \frac{c_k}{\sum_{k'=1}^{V} c_{k'}} \tag{21}$$

For a sufficiently large model that can perfectly fit the training data, the model will predict the probability as $\hat{p}_k$. However, even if the model is confident with this training video, the uncertainty estimation via predictive entropy still gives a non-zero prediction:

$$H[\hat{\mathbf{p}}] = -\sum_k \hat{p}_k \ln \hat{p}_k > 0 \tag{22}$$

It indicates that the uncertainty is incorrectly inflated because there are multiple candidate words as ground-truth next word.

Evidential loss with multiple words acceptable. We use the same notation as above. For evidential learning, the model's output is the evidence e_k of each word k instead of the probability. We first consider the evidential loss without regularization:

$$\begin{aligned} L = -\sum_{n=1}^{N}\sum_{k=1}^{V} y_{nk}[\ln(V + \sum_{k'} e_k) - \ln(e_k + 1)] &= -\sum_{k \in A} c_k[\ln(V + \sum_{k'} e_k) - \ln(e_k + 1)] \\ \text{s.t. } \forall k : e_k >= 0 & \end{aligned} \tag{23}$$

To optimize the objective function, For $k \notin A$:

$$\begin{aligned} &\partial \frac{L}{\partial e_k} = \frac{c_k}{V + \sum_{k'} e_{k'}} > 0 \\ &\hat{e}_k = 0 \end{aligned} \tag{24}$$

It indicates that the optimized evidence for inappropriate words should be zero. Then for $k \in A$:

$$\begin{aligned} &\partial \frac{L}{\partial e_k} = c_k[\frac{1}{V + \sum_{k'} e_{k'}} - \frac{1}{1 + e_k}] < 0 \\ &\hat{e}_k \to \infty \end{aligned} \tag{25}$$

In this case, the vacuity is

$$vac = \frac{V}{V + \sum_{k \in A} e_{k'}} \to 0 \tag{26}$$

Therefore, vacuity is not inflated and indicates that the model is quite confident about the training sample.

Adding regularization usually makes the model generate less confident predictions. However, even with moderate regularization, the vacuity is still small for the training sample. We now consider adding an evidential regularization term to the aggregated loss

$$\begin{aligned} L &= -\sum_{n=1}^{N}\sum_{k=1}^{V} y_{nk}[\ln(V + \sum_{k'} e_k) - \ln(e_k + 1)] + \sum_{n=1}^{N}\sum_{k=1}^{V} \lambda(1 - y_{nk})e_k \\ &= -\sum_{k \in A} c_k[\ln(V + \sum_{k'} e_k) - \ln(e_k + 1)] + \lambda \sum_{k \in A}(N - c_k)e_k + \lambda \sum_{k \notin A} N e_k \\ &\text{s.t. } \forall k : e_k >= 0 \end{aligned} \tag{27}$$

where λ is a weighting parameter. For $k \notin A$:

$$\begin{aligned} &\partial \frac{L}{\partial e_k} = \frac{c_k}{V + \sum_{k'} e_{k'}} + \lambda N > 0 \\ &\hat{e}_k = 0 \end{aligned} \tag{28}$$

It indicates that the optimized evidence for inappropriate words should be zero. Then for $k \in A$:

$$\partial \frac{L}{\partial e_k} = c_k[\frac{1}{V + \sum_{k'} e_{k'}} - \frac{1}{1 + e_k}] + \lambda(N - c_k) \tag{29}$$

We can always find a λ that makes the above derivative negative, and the $\hat{e}_k$ is sufficiently large for $k \in A$. Therefore, the corresponding vacuity is still small.

Table 5: Parameter tuning for evidence-based temperature

Alternative	**UCFCAP**			**MSRVTT**		
	BLEU-4 ↑	ROUGE ↑	CIDEr ↑	BLEU-4 ↑	ROUGE ↑	CIDEr ↑
T=1	0.123	0.335	0.250	0.246	0.484	0.428
T=8	0.115	0.330	0.241	0.247	0.471	0.417
T=4	**0.126**	**0.338**	**0.302**	**0.252**	**0.493**	**0.431**

Remark. With evidential learning, the uncertainty can be further broken into fine-grained components such as vacuity. In some training samples where multiple words can be considered appropriate next word, the model tends to predict relatively high evidence for multiple words. This is captured by dissonance, and not considered for uncertainty quantification. Intuitively, the model should be confident on training samples, and vacuity usually gives us a reasonable uncertainty estimation, as shown in the above proof.

Additional Results

To sample diverse sentences, softmax-based models for predicting word probability typically leverage an adjustable temperature T to modify the logits g (Zhang et al. 2020). However, there are no mechanisms proposed in prior works in the setting of evidential learning. In this paper, we introduce an evidence-based temperature by modifying the opinion of uncertainty from a fixed value 1 to a function based on temperature:

$$p(w_i = k|T) = \frac{e_{i,k} + \exp(T-1)}{\sum_{k'} e_{i,k'} + K\exp(T-1)} \tag{30}$$

This mechanism offers a principled way for sampling, because the predicted evidence, which has probabilistic interpretation, is kept the same without modification. Only the opinion of uncertainty, which is based on the prior knowledge of human, can be adjusted by temperature. With a high temperature, the predicted probability for a specific word $w_i = k$ will approach $1/K$, which is essentially a uniform distribution. In this setting, the hyperparameter T provides us with additional control on the diversity and integrates seamlessly into the evidential setting.

For evidence-based temperature, we leverage the sampling method to generate multiple captions given an input video and apply the temperature to modify the predicted probability based on Eq. (30). We also explore the influence of temperature parameters in the proposed framework, as shown in Table 5. The temperature parameter used for the actual experiment is tuned to $T = 4$. The evidential-based temperature essentially modifies the uncertainty mass without affecting the predicted evidence, and the probability vector for sampling words is dependent on both the uncertainty mass and the evidence of words. A low temperature may result in similar captions generated in multiple sampling rounds and hurt the diversity, while a high temperature may incur too much randomness in caption generation, which also hurts model performance.

We also conduct additional experiment results to illustrate how the tag prediction contributes to the diversity of generated captions. The proposed framework first sample tags based on video features, then sample captions conditioned on tags to encourage the diversity of captions. An alternative approach is directly predicting captions based on video features. To quantitatively measure the diversity, we sample 4 captions per video using the two methods, calculate the BLEU-4 score of each pair of sampled captions, and take the average. Since the temperatures also affect the sampling, we set the temperatures to $T = 2$ and $T = 4$. Intuitively, with a fixed temperature, a low score indicates that the two captions in a pair are dissimilar to each other, and a lower average score indicates the corresponding method generates diverse captions. We provide the averaged score on the test set for the two approaches in Table 6. In addition, some illustrative examples are shown in Figure 8.

Table 6: Comparison of Average Pairwise Similarity

	UCFCAP	**MSRVTT**
No tag (T=2)	0.429	0.371
No tag (T=4)	0.354	0.282
Proposed (T=2)	0.346	0.268
Proposed (T=4)	**0.273**	**0.225**

Additional Discussion on the Datasets and LLM-based Approach

We conduct experiments on two public datasets, the UCFCAP dataset, and MSR-VTT. The UCFCAP dataset is specially chosen because it contains surveillance videos and is in a specialized domain (i.e., security). Active learning is suitable for training

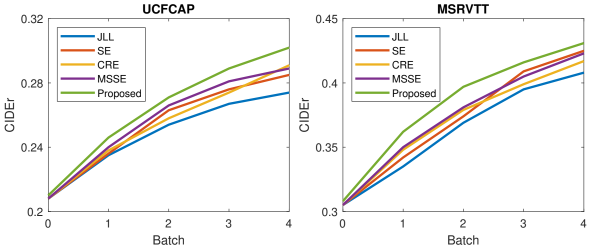
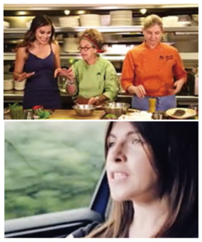


Figure 7: Quantitative comparison on video captioning



Method	Tag	Caption	Method	Caption
Proposed	Show	A television show of cooking	No Tag	A cook in a kitchen is preparing food
	Kitchen	Two woman are preparing food in a kitchen		Women cook in the kitchen
	Cooking	Women cook in the kitchen		Women are preparing food ingredients
	Restaurant	Inside a restaurant kitchen some people are cooking		In a kitchen a woman explaining how to cook
Proposed	Woman	A woman is explaining something	No Tag	A woman drives a car
	Car	A woman is talking in a car		A woman is driving a car
	Recording	Someone is recording while driving a car		A woman is talking and driving a car
	Road	A girl is driving a car on the road		A woman is taking in a car

Figure 8: Illustrative example of predicted captions

models on specialized domains for two major reasons: 1) The data distribution in those domains may differ significantly from common domains (e.g., entertainment and sports). Hence, the off-the-shelf multimodal models pre-trained on crowdsourced data may not perform well in those special domains. 2) In specialized domains, hiring humans to watch videos and annotate captions could be time-consuming and expensive, because it requires domain expertise for annotation. Active learning allows iteratively selecting the most informative samples for labeling, which effectively reduces the cost of annotation and trains the model efficiently.

Nowadays, multimodal foundation models (LLMs) achieved the rapid advancements and demonstrated promising performance. However, general-purpose LLMs may not always provide robust performance, especially when generalizing to specialized domains. They also computationally intensive and suffer from a slower inference speed, limiting their usage in resource-constrained settings (e.g., edge device). In those cases, active learning provides a viable solution.

Limitations and Potential Social Impact

Our work provides a video captioning framework with uncertainty aggregation, and the primary goal is to advance the research in uncertainty estimation and interactive learning. The proposed framework may be applicable to video understanding in specialized domains, such as surveillance, to improve the efficiency of data annotation. However, for general domains where data annotation can be crowd-sourced cost-efficiently, there is less motivation to apply uncertainty estimation for sample selection and human annotation.

To the best of our knowledge, there are no significant ethical issues. However, similar to other machine learning models, inappropriate usage may incur risks and concerns. For example, the model may generate biased or discriminative predictions if the training data is improperly prepared or audited.