Aligning Out-of-Distribution Web Images and Caption Semantics via Evidential Learning

Guohao Sun
Rochester Institute of Technology
Rochester, NY, USA
gs4288@rit.edu

Yue Bai
Northeastern University
Boston, MA, USA
bai.yue@northeastern.edu

Xueying Yang
Santa Clara University
Santa Clara, CA, USA
xyang9@scu.edu

Yi Fang
Santa Clara University
Santa Clara, CA, USA
yfang@scu.edu

Yun Fu
Northeastern University
Boston, MA, USA
yunfu@ece.neu.edu

Zhiqiang Tao
Rochester Institute of Technology
Rochester, NY, USA
zhiqiang.tao@rit.edu

ABSTRACT
Vision-language models, pre-trained on web-scale datasets, have the potential to greatly enhance the intelligence of web applications (e.g., search engines, chatbots, and art tools). Precisely, these models align disparate domains into a co-embedding space, achieving impressive zero-shot performance on multi-modal tasks (e.g., image-text retrieval, VQA). However, existing methods often rely on well-prepared data that less frequently contain noise and variability encountered in real-world scenarios, leading to severe performance drops in handling out-of-distribution (OOD) samples. This work first comprehensively analyzes the performance drop between in-distribution (ID) and OOD retrieval. Based on empirical observations, we introduce a novel approach, Evidential Language-Image Posterior (ELIP), to achieve robust alignment between web images and semantic knowledge across various OOD cases by leveraging evidential uncertainties. The proposed ELIP can be seamlessly integrated into general image-text contrastive learning frameworks, providing an efficient fine-tuning approach without exacerbating the need for additional data. To validate the effectiveness of ELIP, we systematically design a series of OOD cases (e.g., image distortion, spelling errors, and a combination of both) on two benchmark datasets to mimic noisy data in real-world web applications. Our experimental results demonstrate that ELIP improves the performance and robustness of mainstream pre-trained vision-language models facing OOD samples in image-text retrieval tasks. Our implementation is available at https://github.com/heliossun/ELIP.

KEYWORDS
vision-language modeling, evidential learning, uncertainty

1 INTRODUCTION
Web applications [9–11, 13, 19, 35, 45, 47], such as search engines, recommendation systems, etc., greatly benefit daily life, dealing with complicated data formats from different domains (e.g., search engines require massive semantic knowledge, recommendation systems rely on image and text data, etc). Among these web applications, multi-modal data of vision and language (VL) usually play an indispensable role [32], which have attracted remarkable research efforts [41, 43] in recent years. Particularly, CLIP [31] aligns vision and language domains into a shared embedding space, showing a promising zero-shot learning capacity for broad applications. However, web data frequently face many practical challenges, such as low-resolution images due to unreliable internet connections and text marred by garbled characters, leading to numerous out-of-distribution (OOD) samples compared with the clean, well-prepared training data. This gap raises a question – will the pre-trained VL models be vulnerable to OOD samples in web applications?

To investigate the above question, Fig. 1 shows an empirical study of two pre-trained VL models (CLIP [31] and BLIP [20]) for
The performance drop between ID and OOD retrieval of these optimizing an evidential loss. Compared to traditional contrastive
tune the pre-trained VL models to acquire evidence knowledge by
cases. The proposed ELIP develops adapter \cite{12, 15} layers to fine-
tional Language-Image Posterior (ELIP) method, which leverages
evidential learning with VL alignment to improve the generaliza-
development of uncertainty representations based on a single forward pass, enriching
impact of OOD samples on the pre-trained VL models.

Typically, there are three categories of uncertainty modeling: 1) deep ensemble \cite{18}, 2) variational inference \cite{3, 5, 6, 16, 26}, and 3) deep evidential learning \cite{1, 2, 33, 46}. Accounting for the large size of recent VL models, the first two uncertainty estimation methods may be less applicable since they both require multiple inference steps, which can be computationally expensive, especially for the
text ranking problem (where the pairwise calculation occurs). By contrast, deep evidential learning \cite{8} provides explicit uncertainty representations based on a single forward pass, enriching uncertainty knowledge without additional inference costs, which, however, is still under-explored in large-scale VL models.

In this study, we fill in the gap of reasoning uncertainty for VL models by marrying deep evidential uncertainty into a parameter-efficient
tuning framework. Concretely, we propose a novel Evi-
dential Language-Image Posterior (ELIP) method, which leverages evidential learning with VL alignment to improve the generaliza-
tion and reliability of pre-trained VL models in both ID and OOD cases. The proposed ELIP develops adapter \cite{12, 15} layers to fine-
tune the pre-trained VL models to acquire evidence knowledge by optimizing an evidential loss. Compared to traditional contrastive
learning methods that primarily focus on point estimation for the class probability of a sample, the evidential loss considers the entire probability distribution over all samples \cite{33}, improving the model’s robustness against OOD samples and disclosing less confident predictions. Based on the ID and OOD retrieval settings, we conduct extensive experiments to demonstrate the effectiveness of ELIP. Our method outperforms state-of-the-art VL models on image-text retrieval in most OOD cases, showcasing the potential of evidential learning for VL models and its importance in improving model reliability for realistic web scenarios.

2 OUT-OF-DISTRIBUTION SCENARIOS

We introduce two OOD scenarios based on benchmark datasets (e.g., MS-COCO and FLICKR30K), aiming to mimic diverse practical web noisy data to assess the reliability of our model and other mainstream VL models.

Firstly, we introduce simple OOD cases by adding random Gaussian noise to each image with the normal distribution variance as 0.1 or subjecting each image to a random rotation within 0 to 180 degrees. For textual input, we adopt the implementation in \cite{27} to generate naturally noisy text encompassing various error aspects, including diacritics, casing, spelling, suffix/prefix alterations, punctuation variations, whitespace anomalies, word order shifts, insertions, and replacements. Notably, these noisy samples are generated without the reliance on manually designed rules, enhancing the diversity of the perturbations.

Secondly, we introduce web OOD cases (Fig. 2). In realistic web applications, massive amounts of low-quality images are uploaded to the web every day. Some common cases include non-focus images, overexposed images, and compressed images. To mimic such noises, we follow \cite{30} by utilizing blur (zoom), weather (snow), and compression (JPEG) as image-OOD perturbations. Also, the web contains many noisy image descriptions, including spelling and disordered issues. This paper uses word-level synonym replacement (sr) and sentence-level (formal) perturbation to generate noisy captions. We analyze the results aggregated across five perturbation levels for each type of web OOD case. This paper mainly focuses on testing the model’s robustness facing OOD cases. As shown in Fig. 3, we have 10% of simple OOD cases and 90% of web OOD cases over image and text domains. More OOD cases and generation details can be referred to Appendix A.1.
we introduce evidential knowledge to contrastive learning by learning a distribution over the similarities between cross-embeddings. However, Eq. (2) only considers the alignment between correct pairs, without modeling the uncertainty between the query and all the other target samples. To estimate uncertainty in contrastive learning, we introduce evidential knowledge to contrastive learning by learning a distribution over the similarities between cross-embeddings.

Bottleneck Adapter. Adapter module [7, 12, 15] can be easily plugged into existing networks to enable parameter-efficient transfer learning. Specifically, the adapter is a bottleneck structure with linear layers governed by a residual connection between the block’s input and output. This work uses the pre-trained CLIP [31] and BLIP [20] as backbone models. Following [12], we insert one adapter after the self-attention and MLP layers, respectively, in each transformer layer of the vision and language encoders (see Fig. 4). Eventually, the CLIP model has 64M trainable additional parameters, accounting for 13% of the entire model, while the BLIP model has 141M trainable extra parameters, which is 38%. We obtain new image and text features after passing through the pre-trained normalization and linear projection layers. These features are then used to compute the similarities $\rho^{2i}$ and $\rho^{2j}$ in Eq. (1).

3.2 Uncertainty Estimation with Cross Embedding

The evidential deep learning [1, 33, 37] methods overcome the limitations of the standard softmax-based model for uncertainty estimation. Specifically, the softmax function mainly adopts point estimation to quantify the degree of similarity between a given query and multiple targets, which may exhibit low uncertainty in OOD cases. Differently, the evidential deep learning framework models the uncertainty by placing a Dirichlet distribution (Dir) over the prediction probability distribution, allowing to quantify uncertainties under a well-defined theoretical framework based on the Subjective Logic (SL) theory [14]. Typically, SL is beneficial when there are multiple sources of information with varying levels of trustworthiness or when dealing with subjective opinions and beliefs. In this paper, image-text retrieval involves feature alignment and a ranking process that contains multiple sources of information and different levels of trustworthiness, respectively. Therefore, we consider using SL to quantify cross-modal retrieval uncertainty.

The SL reasoning framework generally studies a K mutually exclusive singleton (e.g., class labels) by computing belief mass as

$$b_k = \frac{e_k}{\sum_{i=1}^{K} (e_i + 1)}.$$  (3)

for each singleton $k = 1, \ldots, K$, where $e_k > 0$ denotes the $k^{th}$ singleton’s evidence. Note that the overall uncertainty mass $u$ and all non-negative belief masses are sums up to one, i.e.,

$$u = 1 - \sum_{k=1}^{K} b_k = \frac{K}{\sum_{i=1}^{K} (e_i + 1)}.$$  (4)

The uncertainty $u$ is inversely related to the total amount of evidence $\sum_i e_i$. When there is no evidence for any single entity (each having zero evidence), the aggregate belief equals zero, leading to a maximum uncertainty value of one. Generally, the evidence assigned corresponds to a Dir with parameters $a_k = e_k + 1$. Following [33], given a sample $x_k$ and a classifier $f(\theta)$ with parameters $\theta$, the corresponding Dir has parameters $a_k = f(x_k \mid \theta) + 1$.

This work considers cross-domain information, which differs from the previous methods that use single-domain data. Specifically, we use multi-modal embeddings and define $\alpha$ using the cross-modal similarities between $M$ image-text pairs. Therefore, the subject opinion (belief mass) $b_j^{(i)}$ for the $i^{th}$ query and the $j^{th}$ target sample can be computed from the parameters of the corresponding Dir by

$$b_j^{(i)} = \frac{a_j^{(i)} - 1}{\sum_{l=1}^{M} a_l^{(i)}}.$$  (5)

![Figure 3: Illustration of the percentage of different OOD cases in the image and text domain. We show rotation, Gaussian, and natural for simple OOD, and provide snow, zoom, JPEG, formal, and synonym replacement (sr) for web OOD.](image-url)
Cross-Modal backbone

N×

LN Attn Adapter + LN MLP Adapter + N×

Two giraffes, one looking toward the camera, stand on some grass.

Image-text pair

Uncertainty

Evidence Collection

Evidence - Text Retrieval

Top 5 Text Retrieval:

Cross-embedding similarity as subjective opinions and collects evidence that leads to evidential learning in a single forward pass.

Eventually, Eq. (6) takes input computed in Eq. (1), and the value is greater than zero only for the parallel cross-embeddings. Because $\rho_{ij}$ is computed in the same computation process for image and text, the value is greater than zero only for the parallel $\rho_{ij}$. Because $\alpha$ is used to represent both $\alpha^{it}$ and $\alpha^{ti}$. This work defines the Dir over cross-embeddings between the query and the target samples. By taking the cross-modal similarities $\rho_{ij} \in \mathbb{R}^M$ (the $i^{th}$ row in $\rho$) between the $i^{th}$ query and all the target samples, the $j^{th}$ parameter $a_{ij}^{(i)}$ of the Dir $\alpha^{(i)}$ is computed as

$$a_{ij}^{(i)} = \exp(\rho_{ij}) + 1,$$

where $\rho_{ij}$ represents the $j^{th}$ element in $\rho_{ij}$. We apply $\exp(\cdot)$ as an activation function to ensure positive evidence for all the cross-embeddings. Because $\rho$ is the cross similarity between image and texts, the value is greater than zero only for the parallel pair. Eventually, Eq. (6) takes input computed in Eq. (1), and the output $\alpha$ can be used to calculate uncertainty in Eq. (4). Notably, our proposed $\alpha$ in Eq (6) could connect cross-modal alignment and evidential learning in a single forward pass.

Given $a_{ij}^{(i)}$, our model updates the Dir by using the image-text similarity as subjective opinions and collects evidence that leads to those opinions. During training, the expected matching similarity for the $i^{th}$ query and the $j^{th}$ target sample is computed by

$$E[p_j^{(i)}] = \frac{a_{ij}^{(i)}}{\sum_{i=1}^{M} a_{ij}^{(i)}},$$

where $p_j^{(i)} \in [0, 1]$ indicates the possible values of the probability mass $p$. For convenience, we assign $p$ to represent $p^{it}$ or $p^{ti}$ when no confusion occurs. Throughout the training process, new observations (evidence) would be accumulated to the relevant Dirichlet distribution parameters whenever a query sample corresponds with one of the $M$ target samples. The increment matching similarity between image and text may contribute to its feature alignment, which benefits image/text encoder learning.

### 3.3 Learning with Evidential Knowledge

So far, we have introduced using a Dirichlet distribution to capture evidence knowledge across modalities. In the following, we outline our strategy for fine-tuning the model to optimize the parameters of this distribution. The VL model aims to align two domains into a unified space. Following this concept, we initially compute the cross-embedding similarity. Rather than employing the matching score directly for calculating gradients, we enhance the learning process through two separate steps: 1) gathering model evidence to support correct alignment and 2) minimizing evidence uncertainty when there is poor alignment. Eventually, this allows us to adapt our data to the evidential model at a high level while enforcing a prior to mitigate false evidence and vacuity uncertainty.

**Evidential Loss.** For better clarity, we denote $\alpha^{(i)}$ by $\alpha^{it}/\alpha^{ti}$ as the cross similarities between the $i^{th}$ query and all target samples per image-to-text (i2t) or text-to-image (t2i) retrieval. Given the learned Dirichlet parameters $\alpha^{it}/\alpha^{ti}$, we define evidential
We use dynamic scaling as the following:

\[ \mathcal{L}_{\text{EV}} = \sum_{j=1}^{M} y_j^{\text{ITM}} \left( \psi(a_j^{\text{ITM}}) - \psi(a_j^{\text{TITM}}) \right), \]

\[ \mathcal{L}_{\text{CEV}} = \sum_{j=1}^{M} \frac{y_j^{\text{ITM}}}{y_j^{\text{TITM}}} \left( \psi(a_j^{\text{TITM}}) - \psi(a_j^{\text{ITM}}) \right), \]

where \( \psi() \) is the digamma function and \( s = \sum_{j=1}^{M} \alpha_j \) takes \( a_j^{\text{ITM}} / a_j^{\text{TITM}} \) computed by Eq. (6), denoting \( S^{\text{ITM}} \) or \( S^{\text{TITM}} \), is the Dirichlet strength.

\textbf{Minimizing Evidence on Errors.} The evidential loss aims to align the distribution of image and text features with observed data by optimizing the evidence in favor of the model’s predictions. However, due to the negative samples in the training samples, the model may be misdirected and put strong evidence for the wrong prediction. Thus, we regularize the training by imposing an incorrect evidence penalty, and minimize the evidence of incorrect matching.

We define \( s = y \odot (1 - y) \odot \alpha \), where \( \alpha \) and \( y \) represent \( \alpha^{\text{ITM}} / \alpha^{\text{TITM}} \) and \( y^{\text{ITM}} / y^{\text{TITM}} \). Consequently, we incorporate a Kullback-Leibler (KL) divergence term into the matching loss in (8), where the KL term works as a regularization by penalizing those divergences from negative samples that do not contribute to semantics alignment.

Overall, the evidential loss \( \mathcal{L}_{\text{EO}}(\theta) \) consists of the matching loss and a KL regularization scaled by \( \lambda_t \) as

\[ \mathcal{L}_{\text{EO}} = \mathcal{L}_{\text{ITM}} + \lambda_t KL[D(\phi(\tilde{a})||D(\phi(a)^{\text{TITM}}(1, \cdots, 1)))], \]

\[ \mathcal{L}_{\text{CEV}} = \mathcal{L}_{\text{TITM}} + \lambda_t KL[D(\phi(\tilde{a})||D(\phi(a)^{\text{ITM}}(1, \cdots, 1)))], \]

where \( \lambda_t = \min(1, 15) \) is the annealing coefficient, \( t \) is the index of the current training epoch, \( D(p(1, \cdots, 1)) \) is the uniform Dirichlet distribution, and \( a \) is the Dirichlet parameters of misleading evidence from \( a \). The KL divergence term \( KL[D(p(\tilde{a})||D(\phi(a)^{\text{TITM}}(1, \cdots, 1)))] \) can be computed as

\[ \log \left( \frac{\Gamma(M)}{\Gamma(\sum_{j=1}^{M} \alpha_j)} \right) + \sum_{j=1}^{M} (\alpha_j - 1) [\psi(\tilde{\alpha}_j) - \psi(\tilde{\alpha})]. \]

We use dynamic scaling \( \lambda_t \) to modify the weights of the KL term, leading the model to focus on learning relationships between positive pairs at the beginning and gradually put more attention on negative pairs. Specifically, we enable neural networks to search the parameter space by controlling the impact of the KL divergence, which prevents the network from converging to a uniform distribution for samples that are mis-aligned. Finally, the total loss \( \mathcal{L}_{\text{EO}} \) would update image/text encoders by

\[ \mathcal{L}_{\text{EO}} = \frac{1}{2} (\mathcal{L}_{\text{ITM}} + \mathcal{L}_{\text{CEV}}). \]

We fine-tune the pre-trained CLIP and BLIP models using the evidential loss in (10). By optimizing the inserted adapters, ELIP can preserve high performance on OOD retrieval tasks while achieving reliable performance on OOD retrieval tasks (refer to Table 1). During training with high-level embeddings, the model captures deeper connections between images and text, which enables the generation of evidence for pairwise feature alignment based on these patterns, thereby minimizing the overall loss.

\section{Experiments}

\textbf{Datasets and Evaluation Metrics.} We train and evaluate our model on the MS-COCO [23] and Flickr30K dataset [44]. We evaluate the performance of our model using the common Recall@K (R@K) metric, which measures the proportion of correct matches among the top K retrieved results. Based on different OOD cases across modalities, Table 1 illustrates five evaluation scenarios. Take an example of image retrieval, we report R@K over ID retrieval (T \( \rightarrow \) I), text OOD (T \( \leftrightarrow \) I), image OOD (T \( \rightarrow \) T\( ^* \)), multi-OOD (T \( \rightarrow \) T\( ^* \)), and MultiModal Impact score (MMI) [30] (% of performance drop between ID and OOD retrieval), where the MMI is computed as \( \text{MMI} = (R@K_{ID} - R@K_{OOD}) / R@K_{ID} \). We apply the similar five evaluations for text retrieval in Table 1.

\textbf{Implementation Details.} Our approach is designed to enhance the robustness of pre-trained vision-language models through evidential learning. Therefore, we initialize our implementation by loading the pre-trained CLIP zero-shot and BLIP fine-tuning, namely ELIP and ELIP+, respectively. To fine-tune the model efficiently, we modify the image and text encoder by inserting adapters independently. Specifically, we set the bottle-neck feature dimension to half of the feature dimension from the previous layer, and we use RELU as the activation function. In order to sustain the pre-trained zero-shot performance, we initialize all new parameters of adapters with values drawn from the normal distribution with \( \mu = 0 \) and \( \sigma = 0.001 \). In this work, we leverage deep ensemble [18] to implement an adapter ensemble called CLIP-ensemble, serving as a strong uncertainty-aware baseline. Specifically, CLIP-ensemble freezes the pre-trained CLIP and trains adapters independently with different random seeds. We set the ensemble size as 5 for CLIP-ensemble and took the average prediction during inference. We fine-tuned 15 epochs with a batch size of 200 for ELIP+ and 15 epochs with a 280 batch size for other experiments. We use the AdamW [24] optimizer with an initial learning rate of 5e-5, and the weight decay with a rate of 0.02 for all the experiments.

\subsection{Evaluation on Image-Text Retrieval}

\textbf{MS-COCO.} We split the experimental results into two groups (simple OOD and web OOD). Table 1 provides image-text retrieval and MMI [30] score under simple OOD cases. As can be seen, both ELIP and ELIP+ outperform all baseline models on the MMI benchmark, underscoring the efficacy of our approach in simple OOD scenarios. Despite having fewer trainable parameters, ELIP outperforms previous methods, ALBEF, CLIP, and BLIP, in most OOD retrieval tasks. There is a performance gap between ELIP+ and BLIP among retrieval tasks, which is predictable since ELIP+ is built upon the simplified version of BLIP. Specifically, BLIP incorporates three types of losses: Image-Text Contrastive (ITC), Image-Text Matching (ITM), and Language Modeling (LM). In its ITC component, BLIP utilizes a momentum encoder for soft label generation, enhancing vision-language comprehension and overall model effectiveness. Nonetheless, this encoder is parameter-heavy and requires significant overhead. In contrast, ELIP+ adapts BLIP’s approach but streamlines its structure by omitting momentum encoders, opting for a more efficient fine-tuning method. We also compare ELIP and CLIP-ensemble in terms of performance, robustness, and efficiency. As shown in Table 1, ELIP surpasses CLIP-ensemble on all benchmarks, proving the effectiveness of our method. Also, empirically, ELIP shows less training and inference time than CLIP-ensemble, since it does not require a multi-forward pass over an ensemble.
Table 1: Comparison of performance in terms of Recall@K (R@K) and average MMI score among ID and simple-OOD retrieval on MS-COCO. CLIP and BLIP are pre-trained zero-shot, and the others are fine-tuned on clean MS-COCO. ELIP and ELIP+ are trained based on the pre-trained CLIP and CLIP, respectively.

<table>
<thead>
<tr>
<th>Image Retrieval</th>
<th>T → I</th>
<th>T → I*</th>
<th>T* → I</th>
<th>T* → I*</th>
<th>MMI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@1</td>
<td>R@5</td>
<td>R@10</td>
<td>R@1</td>
<td>R@5</td>
</tr>
<tr>
<td>CLIP [31]</td>
<td>35.3</td>
<td>60.0</td>
<td>70.2</td>
<td>30.4</td>
<td>54.4</td>
</tr>
<tr>
<td>BLIP [20]</td>
<td>56.9</td>
<td>80.8</td>
<td>87.9</td>
<td>43.1</td>
<td>67.8</td>
</tr>
<tr>
<td>ALBEF [21]</td>
<td>60.7</td>
<td>84.3</td>
<td>90.5</td>
<td>47.8</td>
<td>72.0</td>
</tr>
<tr>
<td>BLIP [20]</td>
<td>64.3</td>
<td>85.7</td>
<td>91.5</td>
<td>51.4</td>
<td>74.5</td>
</tr>
<tr>
<td>CLIP-ensemble</td>
<td>58.7</td>
<td>82.8</td>
<td>89.3</td>
<td>50.5</td>
<td>76.3</td>
</tr>
<tr>
<td>ELIP+ (ours)</td>
<td>60.4</td>
<td>83.5</td>
<td>90.2</td>
<td>51.9</td>
<td>76.7</td>
</tr>
<tr>
<td>ELIP+ (ours)</td>
<td>63.7</td>
<td>85.4</td>
<td>91.3</td>
<td>51.0</td>
<td>74.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Text Retrieval</th>
<th>I → T</th>
<th>I* → T</th>
<th>I → T*</th>
<th>I* → T*</th>
<th>MMI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@1</td>
<td>R@5</td>
<td>R@10</td>
<td>R@1</td>
<td>R@5</td>
</tr>
<tr>
<td>CLIP [31]</td>
<td>35.3</td>
<td>60.0</td>
<td>70.2</td>
<td>30.4</td>
<td>54.4</td>
</tr>
<tr>
<td>BLIP [20]</td>
<td>56.9</td>
<td>80.8</td>
<td>87.9</td>
<td>43.1</td>
<td>67.8</td>
</tr>
<tr>
<td>ALBEF [21]</td>
<td>60.7</td>
<td>84.3</td>
<td>90.5</td>
<td>47.8</td>
<td>72.0</td>
</tr>
<tr>
<td>BLIP [20]</td>
<td>64.3</td>
<td>85.7</td>
<td>91.5</td>
<td>51.4</td>
<td>74.5</td>
</tr>
<tr>
<td>CLIP-ensemble</td>
<td>58.7</td>
<td>82.8</td>
<td>89.3</td>
<td>50.5</td>
<td>76.3</td>
</tr>
<tr>
<td>ELIP+ (ours)</td>
<td>60.4</td>
<td>83.5</td>
<td>90.2</td>
<td>51.9</td>
<td>76.7</td>
</tr>
<tr>
<td>ELIP+ (ours)</td>
<td>63.7</td>
<td>85.4</td>
<td>91.3</td>
<td>51.0</td>
<td>74.5</td>
</tr>
</tbody>
</table>

In Table 2, we conduct an analysis of ELIP and other baseline models under web OOD cases. Following [30], we leverage zoom blur, snow noise, and JPEG compression in the vision domain and synonym replacement (sr) and formal (replace normal words with formal words) in the language domain, which are commonly encountered in real-world web applications. The observations reveal that ELIP consistently outperforms all the other baseline models in the context of image-text retrieval in terms of the MMI score. To further test the robustness of ELIP, we provide comparison results on more OOD cases and evaluation metrics in Appendix A.1.

To sum up, we draw our observations on the MS-COCO as follows. 1) Our proposed method improves the robustness of pre-trained models (e.g., CLIP and BLIP) when facing a broad range of OOD cases. 2) We improve the efficiency of finetuning a robust prediction vision-language model, achieving a performance boost, especially compared with existing deep uncertainty methods such as deep ensemble [18]. 3) We found that ELIP can capture reliable similarities between OOD images and OOD text. Specifically, when all inputs are OOD, ELIP can return more accurate retrieval results than ELIP w/o EV based on limited information. However, when image and text are highly damaged without helpful information, the top 1 retrieval will be significantly affected (see Appendix A.1 for detail).

4.1.2 Flickr30k. We further perform our study on the Flickr30K dataset. As shown in Table 3, ELIP outperforms most of the baseline models on simple-OOD retrieval tasks. Also, ELIP and ELIP+ have the smallest performance drop between ID and OOD retrieval. Interestingly, we find that ELIP improves the pre-trained model more than ELIP+. This may be attributed to two factors: 1) the pre-trained CLIP constructs a better cross-embedding than BLIP, and 2) ELIP+ is built upon the simplified version of BLIP, where the model structure, batch size, and query size are minimized to fit our implementation, reducing the model performance empirically. Additionally, the comparison between ELIP and ELIP w/o EV
Table 3: Comparison of performance in terms of Recall@K (R@K) and MMI score among ID and simple-OOD retrieval on Flickr30K. CLIP and BLIP are pre-trained zero-shot, and the others are fine-tuned on clean Flickr30K. ELIP and ELIP+ are transfer learned from pre-trained CLIP and BLIP. I: ID image, T: OOD image, I*: OOD text, and T*: ID text.

<table>
<thead>
<tr>
<th>Image Retrieval</th>
<th>T → I</th>
<th>T → I*</th>
<th>T* → I</th>
<th>T* → I*</th>
<th>MMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLIP [31]</td>
<td>64.5</td>
<td>86.7</td>
<td>92.2</td>
<td>58.0</td>
<td>82.8</td>
</tr>
<tr>
<td>BLIP [20]</td>
<td>78.2</td>
<td>94.0</td>
<td>96.8</td>
<td>61.0</td>
<td>81.2</td>
</tr>
<tr>
<td>ALBEF [21]</td>
<td>85.5</td>
<td>97.5</td>
<td>98.9</td>
<td>68.8</td>
<td>86.6</td>
</tr>
<tr>
<td>BLIP [20]</td>
<td>87.3</td>
<td>97.6</td>
<td>98.9</td>
<td>72.3</td>
<td>89.0</td>
</tr>
<tr>
<td>ELIP w/o EV</td>
<td>85.3</td>
<td>97.9</td>
<td>99.0</td>
<td>78.3</td>
<td>94.3</td>
</tr>
<tr>
<td>ELIP (ours)</td>
<td>86.7</td>
<td>98.0</td>
<td>99.2</td>
<td>78.8</td>
<td>94.4</td>
</tr>
<tr>
<td>ELIP+ (ours)</td>
<td>86.5</td>
<td>97.1</td>
<td>98.3</td>
<td>78.0</td>
<td>94.6</td>
</tr>
</tbody>
</table>

Table 4: Ablation study of the proposed ELIP in terms of average Recall@1 and MMI score in ID and OOD retrieval on MS-COCO. All the models are fine-tuned on MS-COCO.

<table>
<thead>
<tr>
<th>Method</th>
<th>I → T</th>
<th>T → I</th>
<th>I* → T</th>
<th>T* → I*</th>
<th>I → T*</th>
<th>T* → I*</th>
<th>MMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLIP w/o A</td>
<td>60.2</td>
<td>44.5</td>
<td>51.7</td>
<td>38.4</td>
<td>49.8</td>
<td>36.1</td>
<td>43.1</td>
</tr>
<tr>
<td>ELIP w/o IA</td>
<td>71.3</td>
<td>52.8</td>
<td>62.1</td>
<td>45.6</td>
<td>63.8</td>
<td>44.3</td>
<td>55.1</td>
</tr>
<tr>
<td>ELIP w/o TA</td>
<td>76.6</td>
<td>60.1</td>
<td>63.8</td>
<td>51.0</td>
<td>68.0</td>
<td>51.5</td>
<td>55.6</td>
</tr>
<tr>
<td>ELIP w/o EV</td>
<td>76.7</td>
<td>60.3</td>
<td>64.3</td>
<td>51.4</td>
<td>70.5</td>
<td>51.9</td>
<td>58.2</td>
</tr>
<tr>
<td>ELIP (ours)</td>
<td>78.4</td>
<td>60.4</td>
<td>67.2</td>
<td>51.9</td>
<td>72.0</td>
<td>52.3</td>
<td>59.7</td>
</tr>
</tbody>
</table>

Table 5: Domain generalization of image-text retrieval on Flickr30K. All the methods are fine-tuned on MS-COCO.

<table>
<thead>
<tr>
<th>Method</th>
<th>I → T</th>
<th>T → I</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLIP</td>
<td>88.0</td>
<td>98.7</td>
</tr>
<tr>
<td>ALBEF</td>
<td>94.1</td>
<td>99.5</td>
</tr>
<tr>
<td>BLIP</td>
<td>94.8</td>
<td>99.7</td>
</tr>
<tr>
<td>ELIP w/o EV</td>
<td>93.4</td>
<td>99.3</td>
</tr>
<tr>
<td>ELIP</td>
<td>95.2</td>
<td>99.6</td>
</tr>
</tbody>
</table>

As can be seen, ELIP surpasses other methods in most cases and has proved to have good transferability across domains.

### 4.1.4 Limitation Discussion
Although ELIP enables better vision-language modeling for OOD image-text retrieval, it may face the following limitations. 1) **Parameter searching.** The evidential uncertainty is relatively sensitive to the hyperparameter controlling the KL term. This issue might be alleviated by further incorporating hyperparameter optimization techniques [4, 36, 42] or tailoring the activation functions [28] in evidential learning. 2) **Lack of training resources.** The performance of our approach has not been fully optimized due to the lack of sufficient computational resources, e.g., we have not applied large batch sizes and larger pre-trained models.

### 4.2 OOD Detection
ELIP demonstrates an ability to discern between ID and OOD retrieval by using uncertainty as a scoring criterion. As illustrated in Fig. 5, ELIP exhibits a potential to identify anomalous retrieval outcomes when both query and target samples fall in the OOD category. This capability becomes apparent as the estimated uncertainties for OOD image-text retrieval results converge towards a value of 1.0.
following the application of evidential learning. Conversely, during ID retrieval tasks, ELIP consistently furnishes meaningful uncertainty estimations, where the majority of ID retrieval instances yield uncertainties below the threshold of 0.8 since retrieval tasks involve more complex Dirichlet distributions (larger class space) than regular classification tasks, which makes the distribution of uncertainty closer to relative large value. Consequently, ELIP emerges as a reliable uncertainty estimation method, particularly when confronted with OOD problems within the image-text retrieval tasks. It would also be interesting to incorporate recent metric optimization techniques [38, 39] into ELIP toward the OOD setting. Overall, ELIP shows a promising prospect for cross-modal OOD detection.

### 4.3 Ablation Study

In Table 4, we investigate the impact of each component in the proposed ELIP method. By fine-tuning the model on the same data and using the consistent pre-trained weights, we observe that adding image adapters (ELIP w/o TA) has a more considerable improvement than adding text adapters (ELIP w/o IA). This observation implies that the pre-trained vision encoder can extract better semantic knowledge with the assistance of adapters, leading to better cross-modal alignment. Further, the model becomes more robust after training using evidential loss (ELIP) compared to the model fine-tuned without the evidential loss (ELIP w/o EV). When utilizing all components, the effects of adapters and evidential learning complement each other, resulting in substantial improvements compared to regular image-text contrastive learning. Therefore, ELIP achieves the best OOD retrieval and MMI score, improving the robustness of pre-trained models when facing OOD samples.

### 5 RELATED WORK

**Vision-language Modeling and its Web Application.** There are two types of mainstream large-scale vision-language (VL) pre-training models: encoder-based and encoder-decoder structures. Encoder-based methods mainly adopt single-stream or two-stream network architectures. The single-stream uses an individual transformer encoder to concatenate image and text embeddings, e.g., VL-BERT [34], ImageBERT [29], Unified VLP [48], ViLBERT [25], and VisualBERT [22]. In comparison, two-stream methods employ image and text encoders to extract features separately, e.g., CLIP. Some encoder-decoder models leverage cross-modal attention and combine multi-tasks (e.g., image-text retrieval, image captioning) to achieve better performance and higher flexibility on many downstream tasks, e.g., BLIP. In the meantime, due to the demand for large-scale data and the limitation of human-annotated data, most methods use image-data pairs collected from the Web like LAION [32] and VG [17]. In our work, we exploit CLIP, a two-stream approach renowned for its superior image-text alignment capabilities. CLIP represents a significant advancement in creating a flexible and applicable zero-shot classifier; it has a relatively simple structure, with two transformer networks used to extract the image and text features and finally cross-connect during loss calculation. Owing to the impressive zero-shot performance, many works leverage the power of large-scale VL pre-training and benefit the development of web applications [41, 43]. Therefore, large vision-language pre-training plays a significant role in recent web application studies.

### 6 CONCLUSIONS

In this study, we have proposed ELIP to efficiently improve the pre-trained vision-language networks in terms of robustness and performance when handling ID and OOD cases in image-text retrieval tasks via evidence knowledge. Specifically, the proposed ELIP develops cross-domain similarity evidence to approximate the subject opinion of multi-modal alignment during training. Moreover, our method sustains a simple and efficient inference process, making large vision-language models adaptable. ELIP also bridges the gap between evidential learning and fully fine-tuning by leveraging trainable adapters. Our method can easily extend small vision and language encoders to larger ones with more layers. We have provided extensive experimental results encompassing multiple scenarios, catering to ID and OOD image-text retrieval tasks, as well as a detailed ablation study and OOD detection. Particularly, the OOD retrieval covers various noisy settings, including simple noisy and web-style noisy images and texts. Empirical evidence on two public benchmarks has demonstrated the effectiveness of ELIP in facilitating reliable image-text retrieval and precise uncertainty quantification. Our approach’s inherent efficiency and scalability make it particularly valuable for fast and accurate uncertainty estimation in cross-modal retrieval systems. ELIP is especially relevant in fields where safety-critical decisions rely on robust image-text alignment, underscoring the potential impact of our work on broad web applications.
A EXPERIMENTS ON MORE OOD CASES

A.1 OOD Retrieval

Previously, we have provided comparison results of OOD retrieval on 8 OOD cases (see Table 1 and Table 2). To further test ELIP, we generate six more OOD cases (Shot, impulse, speckle, defocus, pixelate, keyboard) based on MS-COCO and provide the comparison results among four methods. To be specific, we have provided brief introductions [30] about the new OOD cases below:

1. **Shot** is an image perturbation characterized by electronic noise, arising from its discrete nature.
2. **Impulse** is an image perturbation that features a color variant of salt-and-pepper noise, which may result from bit errors.
3. **Defocus** is an image perturbation with blur that occurs when an image is out of focus.
4. **Speckle** is an image perturbation, where the noise introduced to a pixel is often more pronounced when the original pixel intensity is higher.
5. **Pixelate** is an image perturbation that occurs when up-sampling a low-resolution image.
6. **Keyboard** is a text perturbation that substitutes character by keyboard distance with a probability p.

Table 6 shows that ELIP improves over the other methods in most of the OOD cases. We also provide qualitative results of cross-domain OOD retrieval in Fig. 6. After generating OOD images and texts based on MS-COCO, we perform ranking and return the top 1 results of ELIP, ELIP w/o EV, and CLIP. As can be seen, ELIP achieves better R@1 results in image and text retrieval tasks.

A.2 Evaluation Metrics

In Table 1, Table 2, and Table 4, we have used the MMI score to measure the performance drop between ID and OOD retrieval. Notably, MMI becomes a valuable supportive metric to gauge the model’s robustness when used with Recall@K. MMI quantifies the impact of perturbations on a model; in other words, it can also describe how sensitive the model is when facing OOD cases.

To provide a more comprehensive evaluation, we employ RSUM (summation of performance) proposed in [40] to evaluate the model’s robustness further, where RSUM is computed as

\[ \text{RSUM} = \sum_i (2\text{R}_{@1} + \text{R}_{@5} + \text{R}_{@10}) + 2\text{R}_{@1} \text{R}_{@5} \text{R}_{@10} \]

Table 7 shows RSUM and average MMI score of all the OOD cases, where ELIP attains the highest average OOD retrieval and the lowest MMI score, demonstrating the robustness of ELIP in handling noisy scenarios. From our observations, while ELIP has a relatively lower RSUM on ID retrieval than ALBEF and BLIP, it presents a higher RSUM in most OOD cases, which indicates the robustness of ELIP when facing noisy images and texts within retrieval tasks. Also, it is predictable that BLIP performs better when dealing with some text OOD cases since they put more effort into improving language understanding.

A.3 OOD Generation

This work introduces simple and web-scaled OOD cases in image-text retrieval tasks. We employ public algorithms to generate OOD samples based on established benchmark datasets (MS-COCO [23] and Flickr30K [44]). This process creates OOD images and texts that simulate real-world noisy data. To produce these noisy images and texts, we apply various perturbation and noising techniques. For basic OOD images, we use two simple yet commonly adopted perturbations: rotation and Gaussian noise. However, recognizing that these simple OOD cases might not accurately represent real-world conditions, we build upon prior research [30] by generating different sets of perturbed images. Specifically, we define five sets of parameters for each perturbation group to adjust the noise level (ranging from 1 to 5), with higher numbers indicating greater noise. In Listing 1, we detail several functions that illustrate our method for generating OOD images and texts.

<table>
<thead>
<tr>
<th>Perturb Method</th>
<th>I → T</th>
<th>T → I</th>
</tr>
</thead>
<tbody>
<tr>
<td>R@1</td>
<td>R@5</td>
<td>R@10</td>
</tr>
<tr>
<td><strong>Shot</strong></td>
<td>CLIP 42.4 69.9 79.9 64.1</td>
<td>34.9 63.3 74.9 57.7</td>
</tr>
<tr>
<td>ALBEF 66.2 86.6 92.0 81.6</td>
<td>52.1 77.9 85.5 71.9</td>
<td></td>
</tr>
<tr>
<td>BLIP 70.1 88.2 82.8 83.7</td>
<td>55.2 79.2 86.5 73.7</td>
<td></td>
</tr>
<tr>
<td>ELIP 71.8 90.1 94.4 85.5</td>
<td>55.7 80.2 88.7 74.6</td>
<td></td>
</tr>
</tbody>
</table>

| Impulse | CLIP 35.6 63.0 74.3 57.6 | 29.8 58.3 70.7 53.0 |
| ALBEF 66.0 86.8 92.1 81.6 | 52.1 77.9 85.8 71.9 |
| BLIP 68.7 87.6 92.3 82.9 | 54.5 78.6 86.1 73.1 |
| ELIP 72.3 90.4 94.7 85.8 | 56.7 81.1 88.5 75.4 |

| Defocus | CLIP 43.7 71.7 81.5 65.6 | 35.2 63.8 75.2 58.1 |
| ALBEF 62.6 84.1 90.1 79.0 | 50.6 75.7 83.9 70.1 |
| BLIP 68.0 87.5 92.2 82.6 | 54.6 78.3 85.4 72.8 |
| ELIP 68.3 89.1 94.2 83.9 | 56.0 80.4 88.0 74.8 |

| Speckle | CLIP 36.5 65.7 77.1 59.8 | 36.5 65.7 77.1 59.8 |
| ALBEF 69.9 89.3 94.1 84.4 | 54.7 80.1 87.6 74.1 |
| BLIP 74.4 91.5 95.0 87.0 | 58.4 81.6 88.5 76.2 |
| ELIP 73.1 91.0 95.1 86.4 | 56.6 81.0 88.3 75.3 |

| Pixel | CLIP 32.4 58.3 68.9 53.2 | 27.3 53.8 65.7 48.9 |
| ALBEF 45.9 65.7 72.7 61.4 | 36.3 58.9 67.5 54.2 |
| BLIP 56.1 76.3 82.6 71.6 | 44.9 68.3 76.5 63.3 |
| ELIP 67.1 88.6 93.4 83.0 | 54.8 79.1 86.9 73.6 |

**Listing 1: Pseudo code for generating OOD samples**

```python
# Simple OOD
# Level1: (0.1, 0)
def Gaussian(X, mean, variance):
    X' = X + gaussian(mean, variance)
    return X'

def Rotation(X):
    X' = rotate(X, angle=random(0,180))
    return X'

# Web OOD
# Level1: (0.1, 0.3, 0.5, 0.8)
def Pixel:
    # Level2: (0.2, 0.3, 0.5, 0.8)
def Keyboard:
    # Level3: (0.55, 0.3, 0.9, 0.8)
# Level4: (0.55, 0.3, 0.5, 0.85, 12, 8, 0.65)
```
Table 7: Comparison of performance in terms of RSUM and MMI score among ID and OOD retrieval. CLIP$_{zs}$ is the pre-trained zero-shot performance, all the other methods are fine-tuned on MS-COCO. OOD$_{\mu}$ is the average RSUM of all OOD retrieval.

<table>
<thead>
<tr>
<th>Method</th>
<th>Clean RSUM</th>
<th>OOD$_{\mu}$ RSUM</th>
<th>Shot RSUM</th>
<th>Impulse RSUM</th>
<th>Speckle RSUM</th>
<th>Defocus RSUM</th>
<th>Pixelate RSUM</th>
<th>Zoom RSUM</th>
<th>Snow RSUM</th>
<th>JPEG RSUM</th>
<th>Keyboard RSUM</th>
<th>SR RSUM</th>
<th>Formal RSUM</th>
<th>MMI RSUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLIP$_{zs}$</td>
<td>394.5</td>
<td>339.1</td>
<td>361.2</td>
<td>330.2</td>
<td>368.7</td>
<td>358.7</td>
<td>308.2</td>
<td>294.6</td>
<td>294.7</td>
<td>388.0</td>
<td>285.5</td>
<td>347.5</td>
<td>393.0</td>
<td>14.0%</td>
</tr>
<tr>
<td>CLIP</td>
<td>420.5</td>
<td>349.8</td>
<td>365.3</td>
<td>331.7</td>
<td>381.5</td>
<td>371.0</td>
<td>306.4</td>
<td>291.0</td>
<td>289.3</td>
<td>402.1</td>
<td>316.1</td>
<td>376.2</td>
<td>417.3</td>
<td>16.8%</td>
</tr>
<tr>
<td>ALBEF</td>
<td>504.6</td>
<td>422.0</td>
<td>460.6</td>
<td>460.3</td>
<td>376.4</td>
<td>447.1</td>
<td>347.0</td>
<td>282.2</td>
<td>408.8</td>
<td>480.9</td>
<td>404.5</td>
<td>471.4</td>
<td>503.1</td>
<td>16.4%</td>
</tr>
<tr>
<td>BLIP</td>
<td>516.6</td>
<td>450.2</td>
<td>472.1</td>
<td>467.7</td>
<td>489.5</td>
<td>466.1</td>
<td>404.7</td>
<td>291.6</td>
<td>432.8</td>
<td>499.6</td>
<td>429.1</td>
<td>484.3</td>
<td>514.4</td>
<td>12.9%</td>
</tr>
<tr>
<td>ELIP</td>
<td>503.5</td>
<td>463.1</td>
<td>480.0</td>
<td>483.7</td>
<td>485.0</td>
<td>476.2</td>
<td>469.8</td>
<td>368.6</td>
<td>448.3</td>
<td>496.9</td>
<td>399.3</td>
<td>484.3</td>
<td>502.4</td>
<td>8.0%</td>
</tr>
</tbody>
</table>

Figure 6: Qualitative results of top 1 cross-domain OOD retrieval (OOD-image: Gaussian noise, random rotate OOD-text: natural noise) on MS-COCO. Left: OOD text retrieval. Right: OOD image retrieval. CLIP$_{zs}$ is zero-shot performance and all the other methods are fine-tuned on MS-COCO.

```python
# Level5: (0.55, 0.3, 2.5, 0.85, 12, 12, 0.55)
def Snow(X, loc, scale, clip, radius, sigma):
    return X + snow_layer(loc, scale, clip, radius, sigma)

# Level1: [1, 1.01, 1.02, ..., 1.11]
def Zoom(X, zoom_factors):
    return (X + zoom(zoom_factors)) / len(zoom_factors)

# Level5: (10, 0.5)
def Defocus(X, radius, alias_blur):
    return defocus(X, kernel(radius, alias_blur))

# Natural text noise
# Natural noise is a mixture of different noisy aspects. To control the noisy level, we sample the error rate of each aspect from a random distribution with different mean value. The default range of mean is (0, 30), where 0 means clean and 30 means all noise. In our project, we set mean to 3.
def Natural_text(X):
    return casing(diacritics(punctuation(spelling(
        whitespace(word-order(wrong
        suffix/prefix(X)))))))
```

# Natural text noise
#Natural noise is a mixture of different noisy aspects. To control the noisy level, we sample the error rate of each aspect from a random distribution with different mean value. The default range of mean is (0, 30), where 0 means clean and 30 means all noise. In our project, we set mean to 3.