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

ACM Reference Format:
Guohao Sun, Yue Bai, Xueying Yang, Yi Fang, Yun Fu, and Zhiqiang Tao. 2024. Aligning Out-of-Distribution Web Images and Caption Semantics via Evidential Learning. In *Proceedings of the ACM Web Conference 2024 (WWW '24), May 13–17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3589334.3645653

WWW '24, May 13–17, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0171-9/24/05.
https://doi.org/10.1145/3589334.3645653

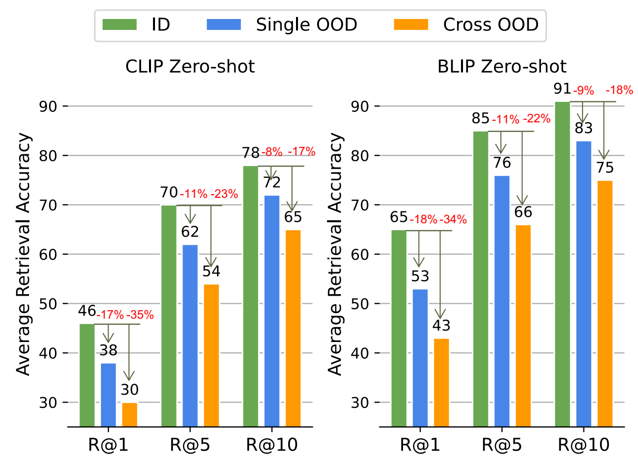

Figure 1: Average performance in terms of Recall@K (R@K) for the image-text retrieval. To measure the vulnerability of large-scale pretraining (e.g., CLIP and BLIP) facing OOD samples, we evaluate them under various noisy cases.

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

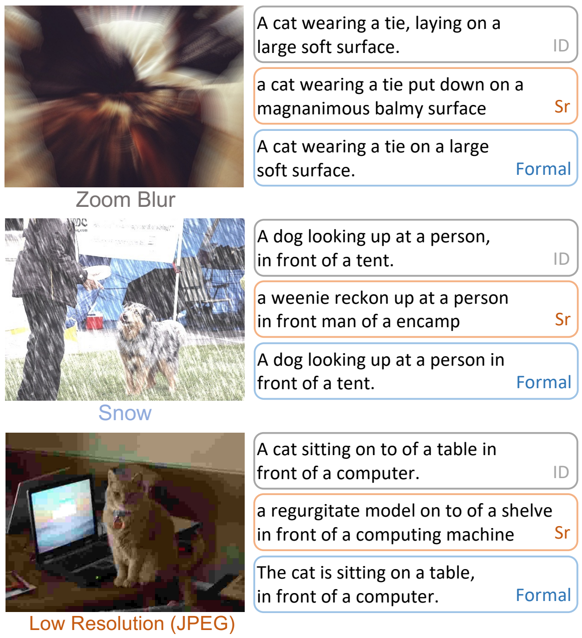

Figure 2: Generated OOD web images and text for cross-modal retrieval. We present web OOD images (e.g., zoom blur, snow, low resolution) paired with one ID and two web OOD texts (e.g., synonym replacement (sr) and formal).

image and text retrieval over in-distribution (ID) and OOD samples. The performance drop between ID and OOD retrieval of these two state-of-the-art models inevitably casts a shadow of directly applying VL models to handle the wild web data. While fine-tuning the VL model with OOD data (varying with different domains) could be a solution, it is highly costly and generally infeasible due to the unknown data on the fly. Thus, we will propose an efficient uncertainty-aware fine-tuning approach to mitigate the negative impact of OOD samples on the pre-trained VL models.

Typically, there are three categories of uncertainty modeling: 1) deep ensemble [18], 2) variational inference [3, 5, 6, 16, 26], and 3) deep evidential learning [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 image-text ranking problem (where the pairwise calculation occurs). By contrast, deep evidential learning [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 Evidential Language-Image Posterior (ELIP) method, which leverages evidential learning with VL alignment to improve the generalization and reliability of pre-trained VL models in both ID and OOD cases. The proposed ELIP develops adapter [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 [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. We summarize the main contributions of this work as follows.

- We introduce and design multiple OOD cases to investigate the vulnerability of large VL models in handling various noises on web data. We provide analysis of the MultiModal Impact (MMI) [30] score and uncertainty estimation based on ID and OOD samples, thoroughly discussing the robustness and reliability of VL models for image-text retrieval tasks.
- We propose a novel uncertainty-aware, parameter-efficient tuning method, namely ELIP. The proposed ELIP adopts evidential learning to integrate image-text matching and uncertainty estimation in a single forward pass.
- Extensive experiments show that our method improves state-of-the-art VL models, CLIP [31] and BLIP [20], in image-text retrieval tasks consisting of diverse OOD samples.

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.

We first introduce ***simple OOD*** cases by adding random Gaussian noise into 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 [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 [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.

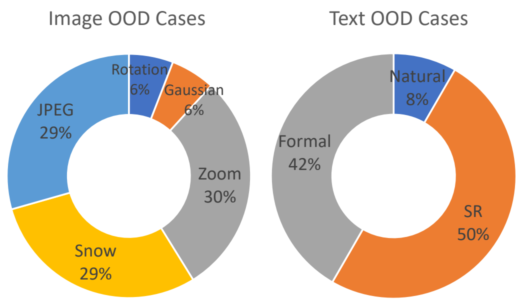

Figure 3: Illustration of the percentage of different OOD cases in the image and text domain. We show rotation, Gaussian, and natural for *simple OOD*, and provide snow, zoom, JPEG, formal, and synonym replacement (sr) for *web OOD*.

3 METHODOLOGY

3.1 Overall Architecture

Vision-language Contrastive Learning. Recent vision-language (VL) models use vision transformers as the image encoder to encode an input image I into a sequence of embeddings as $\{v_{cls}, v_1, \cdots, v_N\}$. They employ a transformer network as the text encoder to project text T into a sequence of embeddings $\{w_{sos}, w_1, \cdots, w_{eos}\}$, where v_{cls} and the activation of the highest layer of the transformer of w_{eos} are treated as extracted features are normalized and linearly projected into a multi-modal D-dimension embedding space. We denote $v \in \mathbb{R}^D$ and $w \in \mathbb{R}^D$ as image and text features, respectively.

Contrastive learning is leveraged to learn a similarity function and capture uni-modal representations. Specifically, the image-to-text and text-to-image similarities between one query sample and M other samples in the target set are computed as

$$\begin{aligned} \rho^{i2t} &= \iota \left\langle v^\top W_0, \cdots, v^\top W_M \right\rangle, \\ \rho^{t2i} &= \iota \left\langle w^\top V_0, \cdots, w^\top V_M \right\rangle, \end{aligned} \tag{1}$$

where ι is a logit-scale, $V \in \mathbb{R}^{M*D}$ and $W \in \mathbb{R}^{M*D}$ are image-text pairs representations, and ρ^{i2t} and ρ^{t2i} can be used to find the correct matching in a top-K list (retrieval), such that parallel image-text pairs should return higher similarity scores. Let y^{i2t} and y^{t2i} be the one-hot label, representing positive samples as 1 and negative samples as 0. The contrastive loss, consisting of image-to-text and text-to-image matching, is defined as

$$\mathcal{L}_{itc} = \frac{1}{2}[\ell(y^{i2t}, \sigma(\rho^{i2t})) + \ell(y^{t2i}, \sigma(\rho^{t2i}))], \tag{2}$$

where σ is a softmax function and $\ell(\cdot,\cdot)$ computes the cross-entropy. 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 cross-alignment, 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 ρ^{i2t} and ρ^{t2i} 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 K mutually exclusive singletons (e.g., class labels) by computing belief mass as

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

for each singleton $k = 1, \cdots, K$, 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.,

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

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 $\alpha_k = e_k + 1$. Following [33], given a sample x_k and a classifier $f(\theta)$ with parameters θ, the corresponding Dir has parameters $\alpha_k = f(x_k \mid \theta) + 1$.

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 α 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

$$b_j^{(i)} = \frac{\alpha_j^{(i)} - 1}{\sum_{l=1}^M (\alpha_l^{(i)})}. \tag{5}$$

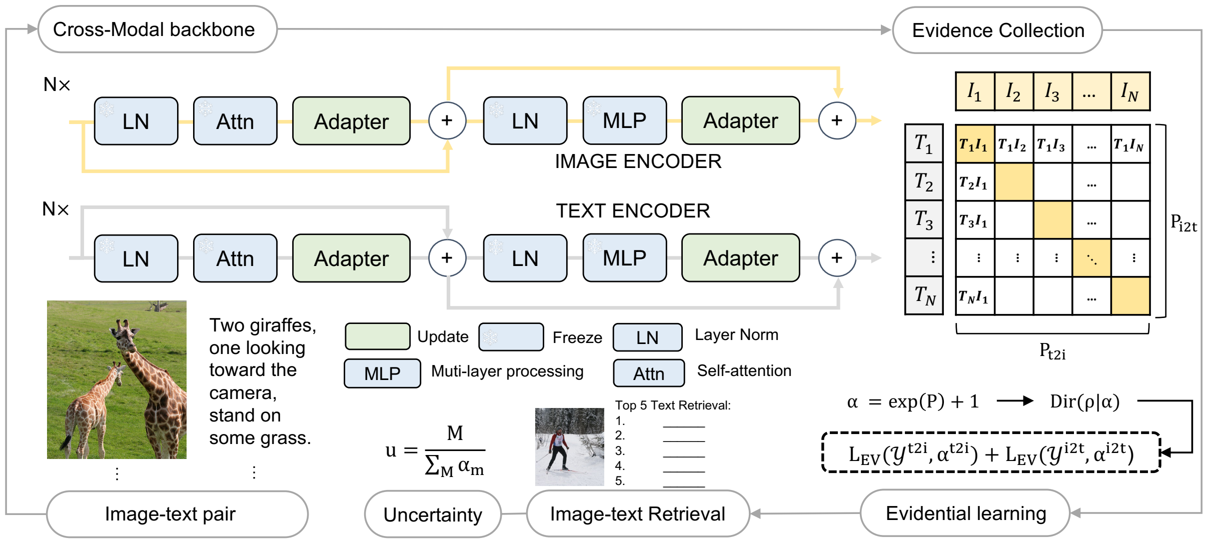

Figure 4: Illustration of the proposed ELIP model. Our method can perform image-text retrieval and uncertainty estimation in a single forward step. The image and text encoders are fine-tuned by different adapters with scalable parameters on clean data without augmentation. We develop a new evidential loss ($\mathcal{L}_{ev}$) to implement image-text matching tasks, and the learned Dirichlet distribution (Dir) posterior is used for uncertainty estimation and OOD detection.

Let $\alpha^{(\mathrm{i})} =< \alpha_1^{(i)}, \cdots, \alpha_M^{(i)} >$ be the parameter of a Dir for cross-modal similarities. Then, we can obtain $(\alpha_j^{(i)} - 1)$ as the evidence estimated by the matching similarity between the i^{th} query and the j^{th} target sample, $1 \leq i, j \leq M$. Upon the obtained parameters, Eq. (4) analytically calculates predictive uncertainties for queries.

To be specific, we define evidence as a measurement of the amount of similarity between the query and target samples in favor of aligning the positive sample and pushing away the negative samples. For convenience, we use the similarity metrics $\rho \in \mathbb{R}^{M*M}$ computed in Eq. (1) to denote ρ^{i2t} or ρ^{t2i}, since image-to-text and text-to-image similarities share the same computation process for evidence. Also, we assign α to represent α^{i2t} and α^{t2i}. This work defines the Dir over cross-embeddings between the query and the target samples. By taking the cross-modal similarities $\rho^{(i)} \in \mathbb{R}^M$ (the i^{th} row in ρ) between the i^{th} query and all the target samples, the j^{th} parameter $\alpha_j^{(i)}$ of the Dir $\alpha^{(i)}$ is computed as

$$\alpha_j^{(i)} = \exp(\rho_j^{(i)}) + 1, \tag{6}$$

where $\rho_j^{(i)}$ represents the j^{th} element in $\rho^{(i)}$. We apply $\exp(\cdot)$ as an activation function to ensure positive evidence for all the cross-embeddings. Because $\rho^{(i)}$ 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 α can be used to calculate uncertainty in Eq. (4). Notably, our proposed α in Eq.(6) could connect cross-modal alignment and evidential learning in a single forward pass.

Given $\alpha_j^{(i)}$, 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

$$\mathbb{E}[p_j^{(i)}] = \frac{\alpha_j^{(i)}}{\sum_{l=1}^{M} \alpha_l^{(i)}}, \tag{7}$$

where $p_j^{(i)} \in [0, 1]$ indicates the possible values of the probability mass p. For convenience, we assign p to represent p^{i2t} orp^{t2i} 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 $\alpha^{i2t}/\alpha^{t2i}$ 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 $\alpha^{i2t}/\alpha^{t2i}$, we define evidential matching losses as the following:

$$\mathcal{L}^{i2t} = \sum_{j=1}^{M} y_j^{i2t}(\psi(S^{i2t}) - \psi(\alpha_j^{i2t})),$$
$$\mathcal{L}^{t2i} = \sum_{j=1}^{M} y_j^{t2i}(\psi(S^{t2i}) - \psi(\alpha_j^{t2i})), \tag{8}$$

where $\psi(\cdot)$ is the *digamma* function and $S = \sum_{j=1}^{M} \alpha_j$ takes $\alpha_j^{i2t}/\alpha_j^{t2i}$ computed by Eq. (6), denoting S^{i2t} or S^{t2i}, is the Dirichlet strength.

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 batch, 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 $\tilde{a} = y + (1-y) \odot a$, where $\tilde{a}$ and y represent $\tilde{a}^{i2t}/\tilde{a}^{t2i}$ and y^{i2t}/y^{t2i}. 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 $\mathcal{L}_{ev}(\theta)$ consists of the matching loss and a KL regularization scaled by λ_t as

$$\mathcal{L}_{ev}^{i2t} = \mathcal{L}^{i2t} + \lambda_t \mathrm{KL}[D(p^{i2t}|\tilde{\alpha}^{i2t})||D(p^{i2t}|\langle 1, \cdots, 1\rangle)],$$
$$\mathcal{L}_{ev}^{t2i} = \mathcal{L}^{t2i} + \lambda_t \mathrm{KL}[D(p^{t2i}|\tilde{\alpha}^{t2i})||D(p^{t2i}|\langle 1, \cdots, 1\rangle)], \tag{9}$$

where $\lambda_t = min(1.0, t/15)$ is the annealing coefficient, t is the index of the current training epoch, $D(p|\langle 1, \cdots, 1\rangle)$ is the uniform Dirichlet distribution, and $\tilde{\alpha}$ is the Dirichlet parameters of misleading evidence from α. The KL divergence term $KL[D(p|\tilde{\alpha})||D(p|\langle 1, \cdots, 1\rangle)]$ can be computed as

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

We use dynamic scaling λ_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 $\mathcal{L}_{EV}$ would evenly update image/text encoders by

$$\mathcal{L}_{EV} = \frac{1}{2}(\mathcal{L}_{ev}^{i2t} + \mathcal{L}_{ev}^{t2i}). \tag{10}$$

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 ID 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.

4 EXPERIMENTS

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 → I), text OOD (T* → I), image OOD (T→ I*), multi-OOD (T* → I*), and MultiModal Impact score (MMI) [30] (% of performance drop between ID and OOD retrieval), where the MMI is computed as $MMI = (R@K_{ID} - R@K_{OOD})/R@K_{ID}$. We apply the similar five evaluations for text retrieval in Table 1.

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 decayed with a rate of 0.02 for all the experiments.

4.1 Evaluation on Image-Text Retrieval

4.1.1 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-short*, and the others are fined-tuned on clean MS-COCO. ELIP and ELIP+ are trained based on the pre-trained CLIP and BLIP, respectively.

Image Retrieval	T → I			T → I*			T* → I			T* → I*			MMI		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP [31]	35.3	60.0	70.2	30.4	54.4	65.3	27.7	50.8	61.3	24.2	46.4	56.9	↓22.3%	↓15.8%	↓12.9%
BLIP [20]	56.9	80.8	87.9	43.1	67.8	76.5	50.0	74.7	82.8	36.9	60.6	70.1	↓23.8%	↓16.2%	↓13.0%
ALBEF [21]	60.7	84.3	90.5	47.8	72.0	80.3	51.9	76.8	85.6	41.2	65.6	74.7	↓22.6%	↓15.2%	↓11.4%
BLIP [20]	**64.3**	**85.7**	**91.5**	51.4	74.5	82.1	**57.2**	**80.3**	**87.4**	45.2	68.8	77.2	↓20.3%	↓13.0%	↓10.1%
CLIP-ensemble	58.7	82.8	89.3	50.5	76.3	**84.7**	50.5	75.8	84.5	42.9	69.0	78.5	↓18.2%	↓11.0%	**↓7.5%**
ELIP (ours)	60.4	83.5	90.2	**51.9**	**76.7**	84.1	52.3	77.0	85.0	44.5	**70.0**	**79.2**	**↓17.9%**	**↓10.7%**	↓8.2%
ELIP+ (ours)	63.7	85.4	91.3	51.0	74.5	82.3	57.0	80.0	87.2	**45.6**	69.3	77.8	↓19.6%	↓12.6%	↓9.7%
Text Retrieval	**I → T**			**I* → T**			**I → T***			**I* → T***			**MMI**		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP [31]	56.0	79.6	86.9	46.3	71.1	79.9	46.1	71.5	80.5	36.6	62.5	73.0	↓23.3%	↓26.7%	↓10.5%
BLIP [20]	72.5	90.0	94.7	52.1	73.4	81.0	67.6	87.9	93.3	48.2	71.1	78.9	↓22.8%	↓13.9%	↓10.9%
ALBEF [21]	77.6	94.3	97.2	59.8	79.5	85.3	71.0	90.6	94.9	54.7	75.7	82.4	↓20.3%	↓13.1%	↓9.9%
BLIP [20]	**81.9**	**95.4**	**97.8**	64.8	82.6	87.6	**76.4**	**93.3**	**96.5**	59.8	79.5	85.5	↓18.2%	↓10.8%	↓8.1%
CLIP-ensemble	76.3	93.2	96.6	65.4	86.1	91.9	69.2	89.6	94.3	59.0	81.8	88.9	↓15.5%	↓7.9%	↓5.1%
ELIP (ours)	78.4	93.6	97.0	**67.2**	**86.4**	**92.0**	72.0	90.6	94.8	59.7	**82.7**	**89.4**	**↓15.4%**	**↓7.5%**	**↓5.1%**
ELIP+ (ours)	81.3	95.2	97.7	64.6	82.6	87.8	76.2	92.9	96.2	**59.9**	79.6	85.4	↓17.7%	↓10.7%	↓8.1%

Table 2: Comparisons of average MMI scores in OOD retrieval. We utilize five web OOD cases generated from MS-COCO, including OOD-image (zoom blur, snow noise, JPEG compression) and OOD-text (synonym replacement (sr), formal).

Image Retrieval	MMI by I^*_{zoom}			MMI by I^*_{snow}			MMI by I^*_{JPEG}			MMI by T^*_{sr}			MMI by T^*_{formal}		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ALBEF [21]	↓51.9%	↓39.1%	↓32.7%	↓26.0%	↓15.8%	↓11.7%	↓8.9%	↓5.1%	↓3.4%	↓13.7%	↓7.8%	↓5.5%	↓0.8%	↓0.5%	**↓0.2%**
BLIP [20]	↓50.5%	↓37.7%	↓31.7%	↓22.7%	↓13.1%	↓9.5%	↓6.5%	↓3.2%	↓2.2%	↓13.7%	↓7.2%	↓5.2%	↓1.2%	↓0.5%	↓0.3%
ELIP (ours)	**↓32.7%**	**↓20.7%**	**↓15.8%**	**↓13.0%**	**↓6.6%**	**↓4.0%**	**↓2.5%**	**↓1.7%**	**↓1.0%**	**↓7.0%**	**↓4.6%**	**↓3.3%**	**↓0.5%**	**↓0.5%**	↓0.4%
Text Retrieval	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ALBEF [21]	↓62.1%	↓45.8%	↓38.1%	↓33.9%	↓18.6%	↓12.8%	↓7.6%	↓3.4%	↓1.9%	↓9.7%	↓3.9%	↓2.2%	↓0.0%	↓0.2%	↓0.2%
BLIP [20]	↓62.5%	↓45.3%	↓37.6%	↓28.8%	↓15.6%	↓10.9%	↓5.4%	↓2.3%	↓1.4%	↓9.4%	↓3.1%	↓1.7%	↓0.2%	**↓0.2%**	↓0.2%
ELIP (ours)	**↓47.9%**	**↓29.4%**	**↓22.5%**	**↓22.2%**	**↓9.9%**	**↓6.1%**	**↓1.4%**	**↓1.2%**	**↓0.6%**	**↓5.8%**	**↓2.7%**	**↓1.1%**	**↓0.0%**	↓0.3%	**↓0.0%**

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 for the experiments on 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-short*, and the others are fined-tuned on clean Flickr30K. ELIP and ELIP+ are transfer learned from pre-trained CLIP and BLIP. I: ID image, I*: OOD image, T: ID text, and T*: OOD text.

Image Retrieval	T → I			T → I*			T* → I			T* → I*			MMI		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP [31]	64.5	86.7	92.2	58.0	82.8	89.2	53.6	79.0	85.6	48.1	74.2	81.5	↓17.5%	↓9.3%	↓7.3%
BLIP [20]	78.2	94.0	96.8	61.0	81.2	87.1	71.3	90.0	93.8	54.6	75.9	82.6	↓20.3%	↓12.4%	↓9.3%
ALBEF [21]	85.5	97.5	98.9	68.8	86.6	91.0	78.6	94.4	96.8	62.4	82.2	87.7	↓18.2%	↓10.0%	↓7.1%
BLIP [20]	**87.3**	97.6	98.9	72.3	89.0	92.8	78.2	94.0	96.8	61.0	81.2	87.1	↓19.2%	↓9.8%	↓6.7%
ELIP w/o EV	85.3	97.9	99.0	78.3	94.3	97.0	78.2	94.2	96.9	70.7	89.1	93.3	**↓11.2%**	↓5.5%	↓3.3%
ELIP (ours)	86.7	**98.0**	**99.2**	**78.8**	94.4	97.0	79.0	**94.5**	**97.1**	**71.1**	**89.9**	**93.9**	↓12.0%	↓5.2%	↓3.2%
ELIP+ (ours)	86.5	97.1	98.3	78.0	**94.6**	**97.4**	**80.4**	94.2	96.3	70.0	89.5	93.3	↓12.0%	**↓4.5%**	**↓2.7%**
Text Retrieval	I → T			I* → T			I → T*			I* → T*			MMI		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP [31]	84.3	97.9	99.3	76.0	93.7	96.6	76.4	94.5	97.5	67.1	90.0	93.7	↓13.2%	↓5.3%	↓3.4%
BLIP [20]	87.4	98.1	99.2	69.1	85.4	89.9	85.8	97.6	98.7	66.4	85.7	89.8	↓15.6%	↓8.7%	↓6.5%
ALBEF [21]	95.9	99.8	100.0	77.2	89.3	91.9	92.4	99.7	99.9	73.9	87.5	90.3	↓15.4%	↓7.6%	↓6.0%
BLIP [20]	**97.2**	99.9	100.0	81.6	92.5	94.8	87.4	98.1	99.2	69.1	85.4	89.9	↓13.0%	↓7.9%	↓5.4%
ELIP w/o EV	96.4	99.8	99.9	88.7	96.6	98.6	91.8	99.5	100.0	84.5	95.4	97.0	↓8.4%	↓2.6%	↓1.4%
ELIP (ours)	95.8	99.8	**100.0**	**88.9**	**97.1**	**98.6**	**94.2**	99.6	**100.0**	**85.9**	**96.1**	**97.9**	**↓6.4%**	**↓2.2%**	**↓1.2%**
ELIP+ (ours)	96.2	**99.9**	**100.0**	87.9	96.8	98.3	93.9	**99.7**	**100.0**	84.3	95.1	96.7	↓7.8%	↓2.7%	↓1.7%

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.

Method	I → T	T → I	I* → T	T → I*	I → T*	T* → I	I* → T*	T* → I*	MMI$_{i2t}$	MMI$_{t2i}$
ELIP w/o A	60.2	44.5	51.7	38.4	49.8	36.1	43.1	30.6	↓19.9%	↓21.3%
ELIP w/o IA	71.3	52.8	62.1	45.6	63.8	44.3	55.1	38.1	↓15.4%	↓19.2%
ELIP w/o TA	76.6	60.1	63.8	51.0	68.0	51.5	55.6	42.3	↓18.5%	↓19.7%
ELIP w/o EV	76.7	60.3	64.3	51.4	70.5	51.9	58.2	43.3	↓16.1%	↓19.0%
ELIP (ours)	**78.4**	**60.4**	**67.2**	**51.9**	**72.0**	**52.3**	**59.7**	**44.5**	**↓15.4%**	**↓17.9%**

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

Method	I → T			T → I		
	R@1	R@5	R@10	R@1	R@5	R@10
CLIP	88.0	98.7	99.4	68.7	90.6	95.2
ALBEF	94.1	99.5	99.7	82.8	96.3	98.1
BLIP	94.8	**99.7**	**100.0**	**84.9**	96.7	98.3
ELIP w/o EV	93.4	99.3	99.7	82.3	96.2	98.2
ELIP	**95.2**	99.6	99.9	83.9	**97.1**	**98.6**

demonstrates the effectiveness of evidential loss, enabling ELIP to achieve more reliable OOD image-text retrieval.

4.1.3 Domain Generalization. To test the transferability across domains, we use the clean Flickr30k as the target domain and treat MS-COCO as the source domain. All the methods (except CLIP) are fine-tuned on the source domain and then tested on the target domain. We report the zero-shot CLIP performance as a baseline. Table 5 summarizes the comparison results on the Flickr30K dataset. 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.

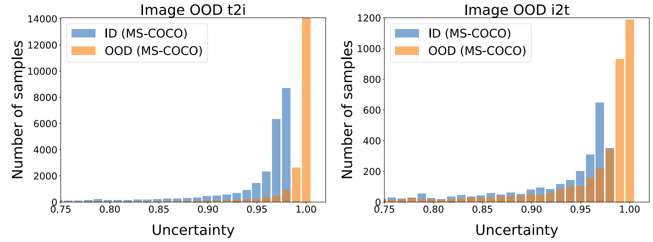

Figure 5: OOD detection by uncertainty of ELIP on ID and OOD image-text retrieval on MS-COCO. The uncertainty values are in the range (0.75–1.00) within each distribution.

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 complemented 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) pretraining 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.

Uncertainty Estimation. Recent studies have shown that uncertainty estimation in DNN contains four different steps [8]: (1) data acquisition, (2) DNN building, (3) applied inference model, and (4) predictive uncertainty model, leading to several factors that may cause model and data uncertainties. Many research efforts have been made to achieve uncertainty estimation. Single deterministic methods predict uncertainty based on the forward pass [1, 2, 33, 46]. Bayesian neural network [3] and its variational approximation [6] have also been applied in modeling weight uncertainty. Plus, the evidential deep learning (EDL) [37] starts to attract attention due to its rich, analytical uncertainty representations and efficient computation. Existing works have applied EDL in uncertainty estimation for both regression [1, 2] and classification tasks [33, 46]. To the best of our knowledge, this study is the first research attempt that incorporates EDL into large vision-language modeling, coping with uncertainty estimation in ODD image/text retrieval tasks.

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.

REFERENCES

[1] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. 2020. Deep evidential regression. In *NeurIPS*.

[2] Wentao Bao, Qi Yu, and Yu Kong. 2021. Evidential Deep Learning for Open Set Action Recognition. In *ICCV*.

[3] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight Uncertainty in Neural Networks. In *ICML*.

[4] Matthias Feurer and Frank Hutter. 2019. *Hyperparameter Optimization.*

[5] Yarin Gal and Zoubin Ghahramani. 2015. Bayesian Convolutional Neural Networks with Bernoulli Approximate Variational Inference. In *ICLR*.

[6] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *ICML*.

[7] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Jiao Qiao. 2021. CLIP-Adapter: Better Vision-Language Models with Feature Adapters. In *ArXiv*.

[8] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseo Lee, Matthias Humt, Jianxiang Feng, Anna M. Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, M. Shahzad, Wen Yang, Richard Bamler, and Xiaoxiang Zhu. 2021. A Survey of Uncertainty in Deep Neural Networks. In *ArXiv*.

[9] Mihajlo Grbovic and Haibin Cheng. 2018. Real-Time Personalization Using Embeddings for Search Ranking at Airbnb. In *SIGKDD*.

[10] U. Gupta, C. Wu, X. Wang, M. Naumov, B. Reagen, D. Brooks, B. Cottel, K. Hazelwood, M. Hempstead, B. Jia, H. S. Lee, A. Malevich, D. Mudigere, M. Smelyanskiy, L. Xiong, and X. Zhang. 2020. The Architectural Implications of Facebook DNN-Based Personalized Recommendation. In *HPCA*.

[11] K. Hazelwood, S. Bird, D. Brooks, S. Chintala, U. Diril, D. Dzhulgakov, M. Fawzy, B. Jia, Y. Jia, A. Kalro, J. Law, K. Lee, J. Lu, P. Noordhuis, M. Smelyanskiy, L. Xiong, and X. Wang. 2018. Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective. In *HPCA*.

[12] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-Efficient Transfer Learning for NLP. In *ICML*.

[13] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning Deep Structured Semantic Models for Web Search Using Clickthrough Data. In *CIKM*.

[14] Audun Jøsang. 2016. *Subjective logic.* Springer.

[15] Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. 2021. Parameter-efficient Multi-task Fine-tuning for Transformers via Shared Hypernetworks. In *ACL*.

[16] Durk P Kingma, Tim Salimans, and Max Welling. 2015. Variational Dropout and the Local Reparameterization Trick. In *NeurIPS*.

[17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2016. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. In *ICCV*.

[18] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2016. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *NeurIPS*.

[19] Chenyi Lei, Shouling Ji, and Zhao Li. 2019. TiSSA: A Time Slice Self-Attention Approach for Modeling Sequential User Behaviors. In *The World Wide Web Conference.*

[20] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*.

[21] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven C. H. Hoi. [n. d.]. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *NeurIPS*.

[22] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2020. VisualBERT: A Simple and Performant Baseline for Vision and Language. In *ACL*, Vol. abs/1908.03557.

[23] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*.

[24] Ilya Loshchilov and Frank Hutter. 2017. Fixing Weight Decay Regularization in Adam. In *ArXiv*.

[25] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *NeurIPS*.

[26] Dmitry Molchanov, Arsenii Ashukha, and Dmitry Vetrov. 2017. Variational dropout sparsifies deep neural networks. In *ICML*.

[27] Jakub N'aplava, Martin Popel, Milan Straka, and Jana Strakov'a. 2021. Understanding Model Robustness to User-generated Noisy Texts. In *W-NUT*.

[28] Deep Shankar Pandey and Qi Yu. 2023. Learn to Accumulate Evidence from All Training Samples: Theory and Practice. In *ICML*.

[29] Di Qi, Lin Su, Jianwei Song, Edward Cui, Taroon Bharti, and Arun Sacheti. 2020. ImageBERT: Cross-modal Pre-training with Large-scale Weak-supervised Image-Text Data. In *ArXiv*.

[30] Jielin Qiu, Yi Zhu, Xingjian Shi, F. Wenzel, Zhiqiang Tang, D. Zhao, Bo Li, and Mu Li. 2022. Are Multimodal Models Robust to Image and Text Perturbations?. In *DMLR*.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*.

[32] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. In *NeurIPS*.

[33] M. Sensoy, Melih Kandemir, and Lance M. Kaplan. 2018. Evidential Deep Learning to Quantify Classification Uncertainty. In *NeurIPS*.

[34] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *ICLR*.

[35] Jiaxi Tang, Francois Belletti, Sagar Jain, Minmin Chen, Alex Beutel, Can Xu, and Ed H. Chi. 2019. Towards Neural Mixture Recommender for Long Range Dependent User Sequences. In *The World Wide Web Conference.*

[36] Zhiqiang Tao, Yaliang Li, Bolin Ding, Ce Zhang, Jingren Zhou, and Yun Fu. 2020. Learning to Mutate with Hypergradient Guided Population. In *NeruIPS*.

[37] Dennis Ulmer. 2021. A survey on evidential deep learning for single-pass uncertainty estimation. In *arXiv*.

[38] Zitai Wang, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. 2022. OpenAUC: Towards AUC-Oriented Open-Set Recognition. In *NeruIPS*.

[39] Zitai Wang, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. 2023. Optimizing Partial Area Under the Top-k Curve: Theory and Practice. *TPAMI* (2023).

[40] Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. 2019. Unified Visual-Semantic Embeddings: Bridging Vision and Language With Structured Meaning Representations. *CVPR*.

[41] Jheng-Hong Yang, Carlos Lassance, Rafael Sampaio de Rezende, Krishna Srinivasan, Miriam Redi, Stéphane Clinchant, and Jimmy Lin. 2023. AToMiC: An Image/Text Retrieval Test Collection to Support Multimedia Content Creation. In *SIGIR*.

[42] Xueying Yang, Jiamian Wang, Xujiang Zhao, Sheng Li, and Zhiqiang Tao. 2022. Calibrate Automated Graph Neural Network via Hyperparameter Uncertainty. In *CIKM*.

[43] Linli Yao, Wei Chen, and Qin Jin. 2022. CapEnrich: Enriching Caption Semantics for Web Images via Cross-modal Pre-trained Knowledge. In *Proceedings of the ACM Web Conference 2023.*

[44] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In *TACL*.

[45] Chengliang Zhang, Minchen Yu, Wei Wang, and Feng Yan. 2019. MArk: exploiting cloud services for cost-effective, SLO-aware machine learning inference serving. In *USENIX Conference on Usenix Annual Technical Conference.*

[46] Xujiang Zhao, Feng Chen, Shu Hu, and Jin-Hee Cho. 2020. Uncertainty aware semi-supervised learning on graph data. In *NeurIPS*.

[47] Kaifu Zheng, Lu Wang, Yu Li, Xusong Chen, Hu Liu, Jing Lu, Xiwei Zhao, Changping Peng, Zhangang Lin, and Jingping Shao. 2022. Implicit User Awareness Modeling via Candidate Items for CTR Prediction in Search Ads. In *ACM Web Conference.*

[48] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. 2020. Unified Vision-Language Pre-Training for Image Captioning and VQA. In *AAAI*.

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 upsampling 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.

Table 6: Comparison of performance in terms of Recall@k in OOD retrieval on MS-COCO. All the methods are fine-tuned on MS-COCO.

Perturb	Method	$I^* \rightarrow T$				$T \rightarrow I^*$			
		R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean
Shot	CLIP	42.4	69.9	79.9	64.1	34.9	63.3	74.9	57.7
	ALBEF	66.2	86.6	92.0	81.6	52.1	77.9	85.5	71.9
	BLIP	70.1	88.2	82.8	83.7	55.2	79.2	86.5	73.7
	ELIP	**71.8**	**90.1**	**94.4**	**85.5**	**55.7**	**80.2**	**87.7**	**74.6**
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**
		$I \rightarrow T^*$				$T^* \rightarrow I$			
Keyboard	CLIP	36.8	62.1	72.8	57.2	21.0	41.2	51.6	37.9
	ALBEF	57.9	82.6	89.6	76.7	38.0	63.4	73.0	58.1
	BLIP	**64.1**	**86.4**	**91.9**	**80.8**	**42.7**	**67.5**	**76.6**	**62.3**
	ELIP	58.2	82.5	89.5	76.7	36.8	61.3	71.0	56.4

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

$$RSUM = SUM(i2t(R@1, R@5, R@10) + t2i(R@1, R@5, R@10)).$$

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.

Listing 1: Pseudo code for generating OOD samples

```
# 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: (0.1,0.3,3,0.5,10,4,0.8)
# Level2: (0.2,0.3,2,0.5,12,4,0.7)
# Level3: (0.55,0.3,4,0.9,12,8,0.7)
# Level4: (0.55,0.3,4.5,0.85,12,8,0.65)
```

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_{μ} is the average RSUM of all OOD retrieval.

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

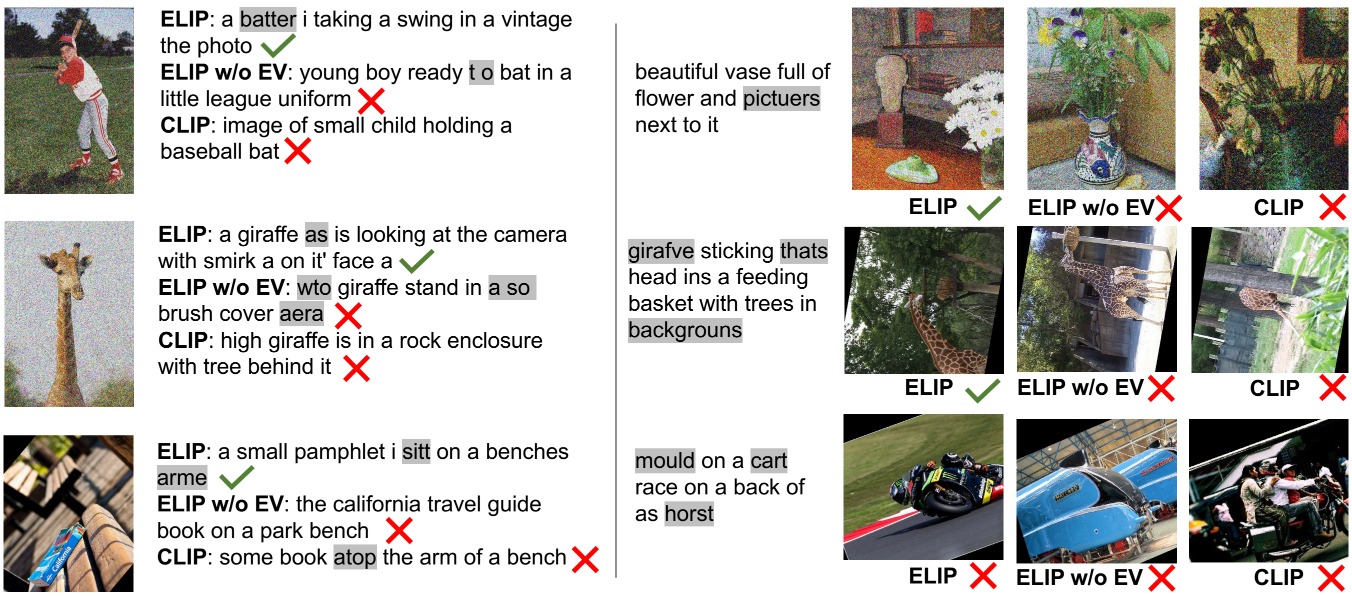

Figure 6: Qualitative results of top 1 cross-domain OOD retrieval (OOD-image: Gaussian noise, random rotate OOD-text: natural noise) on MS-COCO. Left: OOD text retrieval. Right: OOD image retrieval. $CLIP_{zs}$ is *zero-shot* performance and all the other methods are fine-tuned on MS-COCO.

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

# Level1: [1, 1.01, 1.02, ..., 1.11]
# Level2: [1, 1.01, 1.02, ..., 1.16]
# Level3: [1, 1.02, 1.04, ..., 1.21]
# Level4: [1, 1.02, 1.04, ..., 1.26]
# Level5: [1, 1.03, 1.06, ..., 1.33]
def Zoom(X, zoom factors):
    X' = (X + zoom(zoom factors)) / len(zoom factors)
    return X'

# Level1: (3, 0.1)
# Level2: (4, 0.5)
# Level3: (6, 0.5)
# Level4: (8, 0.5)
# Level5: (10, 0.5)
def Defocus(X, radius, alias_blur):
    X' = defocus(X, kernel(radius, alias_blur))
    return 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.
def Natural_text(X):
    X' = casing(diacritics(punctuation(spelling(
           whitespace(word-order(wrong
                suffix/prefix(X)))))))
    return X'
```