Appendix A Proof of Theorem 1

First, we define our id label is the y, which means the one-hot distribution for groudtruth and our ood label is the $\mathcal{U}$, which is the uniform distribution. Like the related work, we follow the setup of them and define our id label smooth are as follow:

$$\widetilde{y} = u \cdot f_g(x) + (1 - u) \cdot \mathcal{U} \tag{1}$$

Consider ℓ to be the convex logistic loss function applied to binary classification tasks. And considering the property of convex function, we have:

$$\ell\Bigg(f_\theta(x), \widetilde{y}\Bigg) = \ell\Bigg(f_\theta(x), u \cdot f_g(x) + (1-u) \cdot \mathcal{U}\Bigg) \leq (1-u) \cdot \ell\Bigg(\mathcal{U}, f_\theta(x)\Bigg) + u \cdot \ell\Bigg(f_g(x), f_\theta(x)\Bigg) \tag{2}$$

According to our definition of generalisation error, we have the following:

$$\begin{aligned}
GE(f, \widetilde{y}) &= \mathbb{E}_{(x,y)\sim D_{OOD}} \ell\Bigg(f_\theta(x), \widetilde{y}\Bigg) \\
&= \mathbb{E}_{(x,y)\sim D_{OOD}} \ell\Bigg(f_\theta(x), (1-u) \cdot \mathcal{U} + u \cdot f_g(x)\Bigg) \\
&\leq \mathbb{E}_{(x,y)\sim D_{OOD}} \Bigg[(1-u) \cdot \ell(\mathcal{U}, f_\theta(x)) + u \cdot \ell(f_g(x), f_\theta(x))\Bigg] \\
&= \mathbb{E}_{(x,y)\sim D_{OOD}} \Bigg[\ell(\mathcal{U}, f_\theta(x))\Bigg] - \mathbb{E}_{(x,y)\sim D_{OOD}} \Bigg[u \cdot \ell\,(\mathcal{U}, f_\theta(x))\Bigg] + \mathbb{E}_{(x,y)\sim D_{OOD}} \Bigg[u \cdot \ell(f_g(x), f_\theta(x))\Bigg] \\
&= \Bigg(1 - \mathbb{E}_{(x,y)\sim D_{OOD}}[u]\Bigg) \cdot \mathbb{E}_{(x,y)\sim D_{OOD}} \Bigg[\ell(\mathcal{U}, f_\theta(x))\Bigg] - Cov\Bigg(u, \ell(\mathcal{U}, f_\theta(x))\Bigg) \\
&\quad + \mathbb{E}_{(x,y)\sim D_{OOD}}\,[u] \cdot \mathbb{E}_{(x,y)\sim D_{OOD}} \Bigg[\ell(f_g(x), f_\theta(x))\Bigg] + Cov\Bigg(u, \ell(f_g(x), f_\theta(x))\Bigg) \\
&\leq Cov\Bigg[u, \ell(f_g(x), f_\theta(x))\Bigg] - Cov\Bigg[u, \ell(\mathcal{U}, f_\theta(x))\Bigg] \\
&\quad + \underbrace{\mathbb{E}_{(x,y)\sim D_{OOD}} \Bigg[\ell(\mathcal{U}, f_\theta(x))\Bigg] + \mathbb{E}_{(x,y)\sim D_{OOD}} \Bigg[\ell(f_g(x), f_\theta(x))\Bigg]}_{constant} \\
&= Cov\Bigg[u, \ell(f_g(x), f_\theta(x))\Bigg] - Cov\Bigg[u, \ell(\mathcal{U}, f_\theta(x))\Bigg] + C
\end{aligned} \tag{3}$$

Among them, the last two items are defined as irrelevant items C that are irrelevant to u. In addition, in many research works (Yuan et al., 2020), the relationship between soft labels and distillation learning is explored. It is believed that by using soft labels and, the loss corresponding to distillation learning can be reduced, that is, C converges to an empirical error, which can also be considered a constant.

Within our theoretical setup and under the stated assumptions, reducing the generalization error bound requires satisfying the conditions that $Cov[u^*, KL(f_g(x), f_\theta(x))] < 0$ and $Cov[u^*, KL(\mathcal{U}, f_\theta(x))] > 0$. This leads to two corollaries for the design of u that it must be negatively correlated with the $KL(f_g(x), f_\theta(x))$, and positively correlated with the $KL(\mathcal{U}, f_\theta(x))$.

Appendix B OOD Score

The CLIP model's multimodal feature alignment capability enables the MCM Ming et al. (2022) method to perform zero-shot OOD detection by quantifying the similarity distribution between image features and C class text embeddings. The OOD Score function is defined as follows:

$$S_{_{MCM}} = \max_i \frac{\exp\left(\langle \phi_I(\mathbf{x}), \phi_T(\mathbf{t}_i) \rangle / \tau\right)}{\sum_{j=1}^{C} \exp\left(\langle \phi_I(\mathbf{x}), \phi_T(\mathbf{t}_j) \rangle / \tau\right)} \tag{4}$$

where $\tau = 1$ is the temperature parameter, and $\langle \cdot, \cdot \rangle$ denotes cosine similarity.

By introducing a global-local hierarchical feature matching mechanism, GL-MCM Miyai et al. (2025) extends the OOD score calculation to:

$$S_{_{GL-MCM}} = \max_i \frac{\exp\left(\langle \phi_I(\mathbf{x}^{local}), \phi_T(\mathbf{t}_i) \rangle / \tau\right)}{\sum_{j=1}^{C} \exp\left(\langle \phi_I(\mathbf{x}^{local}), \phi_T(\mathbf{t}_i) \rangle / \tau\right)} + S_{_{MCM}} \tag{5}$$

where $\mathbf{x}^{local}$ represents the feature of the i-th local image patch.

Appendix C Experimental Details

Base OOD Benchbark. The implementation of the system adheres to the LoCoOp framework with CLIP-ViT-B/16 Dosovitskiy et al. (2020), where the feature maps exhibit a spatial resolution of 14x14. The key hyperparameters have been empirically configured as follows: the neighbourhood size K = 200 across all experiments, the knowledge distillation coefficient α = 0.25, and the regularization weight λ = 0.3. The additional training specifications encompass 50 epochs with a base learning rate of 0.002, a batch size of 32, and a prompt token length of N=16. It is imperative that all experiments are conducted on a single NVIDIA A6000 GPU in order to ensure hardware consistency.

Hard OOD Benchbark. It is evident that our fundamental experimental details are consistent with those of the baseood benchmark. However, given that imagenet-10 and imagenet-20 contain 10 and 20 data types respectively, it was determined that the neighborhood size K=2 would be employed for these hard-to-imitate experiments. The results of the model under the 16-shot setting are presented in full in our paper.

OpenOOD OOD Benchbark. The experimental details are fundamentally analogous to the base food benchmark. The imagenet1k has been selected as the ID dataset, while the SSh-hard, NINCO and OpenImage-O have been designated as the OOD dataset. It should be noted that iNaturalist and Texture have not been included in the evaluation process, as these two datasets have previously been evaluated in the base OOD benchmark.

Appendix D The Selection of A Suitable General Knowledge Model

Table 1: The cross-domain generalisation performance of prompt-tuned general knowledge models f_g, pre-trained on ImageNet-21K and evaluated through out-of-distribution benchmarks.

	OOD Dataset									
Method	iNaturalist		SUN		Places		Texture		Average	
	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC
					MCM					
$\texttt{LUS}_{\texttt{CLIP}}$	**27.74**	94.16	34.78	93.01	42.55	90.19	48.48	89.05	38.39	91.60
$\texttt{LUS}_{\texttt{POMP}}$	30.80	**94.17**	**31.25**	**93.91**	**39.78**	**90.79**	**41.50**	**90.81**	**35.83**	**92.42**
					GL-MCM					
$\texttt{LUS}_{\texttt{CLIP}}$	**13.59**	**96.81**	27.73	93.87	35.94	91.09	51.21	85.80	32.12	91.89
$\texttt{LUS}_{\texttt{POMP}}$	16.41	96.48	**22.78**	**95.05**	**32.41**	**91.80**	**44.11**	**88.95**	**28.92**	**93.07**

The following experiments are presented, in which other models of general knowledge are selected to guide the model in acquiring general knowledge. The POMP paper Ren et al. (2023) was selected as the secondary general knowledge model to present the experimental results. POMP presented the results of prompt tuning on the ImageNet-21K dataset. In this instance, the model under discussion was employed. It is evident that the parameter settings are consistent with the base OOD benchmark. Our results are shown in Table 1, where the clip subscript represents our general knowledge as " a photo of ", and the POMP subscript represents this general knowledge after training on Imagenet-21k. Our results demonstrate that different $f_g(x)$ models can exhibit varying performance for our method, indicating that our model will acquire distinct general knowledge under distinct $f_g(x)$ settings.

Moreover, in order to demonstrate the rationality of our methodology, we employ the same comparison strategy as outlined in Table 1. The results of the ood score of POMP using MCM and GL-MCM in ood detection are presented, as well as the results of the ood score of the LoCoOp model using only our training loss. The following presentation will outline the output results of the model under the KDE strategy. The results of the study are presented in tabular form. The findings of this study suggest that the proposed methodology explores the upper limit of OOD detection, while exhibiting the POMP generalization.

Table 2: The model performance of POMP when used as the f_g model. The present method has been developed in such a manner that it inherits the generalisation ability of POMP, whilst also exploring the upper limit of OOD detection.

Method	OOD Dataset									
	iNaturalist		SUN		Places		Texture		Average	
	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC
MCM										
LoCoOp	38.96	92.34	32.40	93.60	37.95	**91.00**	49.32	88.70	39.65	91.41
LUS	**30.80**	**94.17**	**31.25**	**93.91**	**39.78**	90.79	41.50	**90.81**	**35.83**	**92.42**
GL-MCM										
LoCoOp	24.38	94.95	25.45	94.77	32.63	**91.81**	52.32	86.58	33.69	92.03
LUS	**16.41**	96.48	**22.78**	**95.05**	**32.41**	91.80	**44.11**	**88.95**	**28.92**	**93.07**

Appendix E More Experimental Results

The appendices to this section contain further experimental results of our model, the purpose of which is to demonstrate its experimental performance. The following presentation comprises the experimental results of MCM and GL-MCM under a variety of conditions.

Table 3: cross-domain OOD detection performance comparison across OOD datasets which under different detection frameworks setting: evaluations follow the OpenOOD benchmark with ImageNet-1K as ID data against SSB-hard, NINCO, and OpenImage-O OOD splits, and the MCM cross-evaluation protocol adopting ImageNet-10 ImageNet-20 as ID datasets with reciprocal OOD testing . Our first row represents the id dataset and the second row represents the ood dataset.

Method	**ImageNet-10**		**ImageNet-20**		**ImageNet-1K**						**Average**	
	ImageNet-20		**ImageNet-10**		**SSh-hard**		**NINCO**		**OpenImage-O**			
	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC
LoCoOp	28.20	92.75	34.40	92.34	90.27	63.16	82.54	69.19	45.12	90.73	56.11	81.63
Ours	5.70	98.60	16.10	97.66	88.78	64.41	79.19	74.10	41.43	91.84	46.24	85.32

The experimental results obtained under the OpenOOD and MCM benchmarks demonstrate that GL-MCM exhibits superior performance in cross-dataset ID and OOD detection scenarios when compared to the baseline.

The experimental findings yielded from the execution of MCM benchmarks demonstrate that GL-MCM evinces superior performance in OOD detection scenarios when contrasted with the baseline MCM. This outcome is congruent with our experimental expectations and concomitantly signifies that GL-MCM also attains comparatively favourable enhancement results for GL-MCM of our soft label.

Table 4: OOD detection performance for ImageNet-1k as ID, the SSh-hard, NINCO, OpenImage-O as OOD dataset.

Method	**ImageNet-1K**					
	SSh-hard		NINCO		OpenImage-O	
	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC
$\texttt{LUS}_{\texttt{MCM}}$	88.78	64.41	79.19	74.10	41.43	91.84
$\texttt{LUS}_{\texttt{GL}}$	85.13	68.27	72.57	76.06	34.59	92.36

Table 5: OOD detection performance for ImageNet-10, ImageNet-20 as ID, the corresponding imagenet20, imagenet10 as ood datasetas.

Method	ImageNet10 ImageNet20		ImageNet20 ImageNet10	
	FPR95	AUROC	FPR95	AUROC
$\texttt{LUS}_{\texttt{MCM}}$	5.70	98.60	16.10	97.66
$\texttt{LUS}_{\texttt{GL}}$	10.60	98.66	9.90	98.32

The subsequent presentation will expound upon the findings of the model's image detection process in relation to imaget100, which will be utilised as the ID data. The experimental results of the model on 4-shot are also presented. In the present experiment, the value of K was set to 20. The 1-shot configuration was not selected as the experimental outcome due to the inability of our model to converge on the original LoCoOp setting. In order to conduct a one-shot experiment, it is necessary to enlarge the epoch under the LoCoOp setting until the experimental results obtained are consistent with those reported in the aforementioned paper. The present study employs imagenet-100 as the ID dataset, thereby adopting a methodology that explores enhanced object detection while ensuring the efficacy of the $f_g(x)$ model. This approach is employed to demonstrate the efficacy of the proposed methodology.

Appendix F Comparing with more uncertainty method.

Static weight. We first define the static method which use the weight is 1/2. We define the soft label for OOD data as follows:

$$\widetilde{y} = \frac{1}{2} \cdot f_g(x) + \frac{1}{2} \cdot \mathcal{U} \tag{6}$$

Max logit. We initially define the uncertainty measure as the maximum logit, denoted as:

$$u = \max_{c \in \mathcal{C}} f_c(\mathbf{x}) \tag{7}$$

where $f_c(\mathbf{x})$ is the logit output for class c given input $\mathbf{x}$, and $\mathcal{C}$ is the set of all classes.

Since this raw uncertainty value is not normalized, we scale it to the range $[0, 1]$ using extremal statistics from the entire training dataset $\mathcal{D}_{\text{train}}$. Let:

$$u_{\min} = \min_{\mathbf{x}_i \in \mathcal{D}_{\text{train}}} \max_{c} f_c(\mathbf{x}_i) \tag{8}$$

$$u_{\max} = \max_{\mathbf{x}_i \in \mathcal{D}_{\text{train}}} \max_{c} f_c(\mathbf{x}_i) \tag{9}$$

represent the global minimum and maximum uncertainty values observed over $\mathcal{D}_{\text{train}}$. The normalized uncertainty u_{norm} is then defined as:

$$u_{\text{norm}} = \frac{u - u_{\min}}{u_{\max} - u_{\min}} \tag{10}$$

This min-max normalization ensures $u_{\text{norm}} \in [0, 1]$ with the property that the most uncertain sample in the training set maps to 1 and the least uncertain to 0.

$$u = \frac{u - u_{min}}{u_{max} - u_{min}} \tag{11}$$

Table 6: OOD detection performance for ImageNet-10, ImageNet-20 as ID, the corresponding imagenet20, imagenet10 as ood datasetas.

Method	ImageNet10 ImageNet20		ImageNet20 ImageNet10	
	FPR95	AUROC	FPR95	AUROC
$\texttt{LUS}_{\texttt{MCM}}$	5.70	98.60	16.10	97.66
$\texttt{LUS}_{\texttt{GL}}$	10.60	98.66	9.90	98.32

Table 7: Cross-domain generalization performance on ImageNet-100 as ID data under four-shot learning protocol. A comparison was made between MCM and LoCoOp.

Method	OOD Dataset									
	iNaturalist		SUN		Places		Texture		Average	
	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC
					MCM					
$\text{LoCoOp}_{\text{MCM}}$	18.69	96.54	21.16	96.32	27.82	95.12	26.17	94.99	23.46	95.74
$\texttt{LUS}_{\texttt{MCM}}$	**10.70**	**97.71**	**16.81**	**96.92**	**22.52**	**95.65**	**24.68**	**95.49**	**18.67**	**96.44**
					GL-MCM					
$\text{LoCoOp}_{\text{GL}}$	12.97	97.09	**12.55**	97.20	**18.15**	96.06	**26.17**	94.36	17.46	96.18
$\texttt{LUS}_{\texttt{GL}}$	4.44	98.87	13.15	**97.42**	18.43	96.11	27.23	**94.48**	**15.81**	**96.72**

Entropy. The entropy-based uncertainty is defined as $u = -\sum_c p_c(\mathbf{x}) \log p_c(\mathbf{x})$ and normalized to [0,1] using:

$$u_{\text{norm}} = \frac{u - u_{\min}}{u_{\max} - u_{\min}} \tag{12}$$

where $u_{\min}$ and $u_{\max}$ are the extreme entropy values from the training set.

Appendix G More Temperature Coefficient Visualization Results.

This section analyzes the convergence of u under different hyperparameter settings in the paper. These images match our analysis in the article. For smaller temperature coefficients, u will have large fluctuations, while for larger temperature coefficients, the fluctuations are smaller, but the performance deteriorates. In the experiments in the paper, we choose the results when the temperature coefficient is 1.

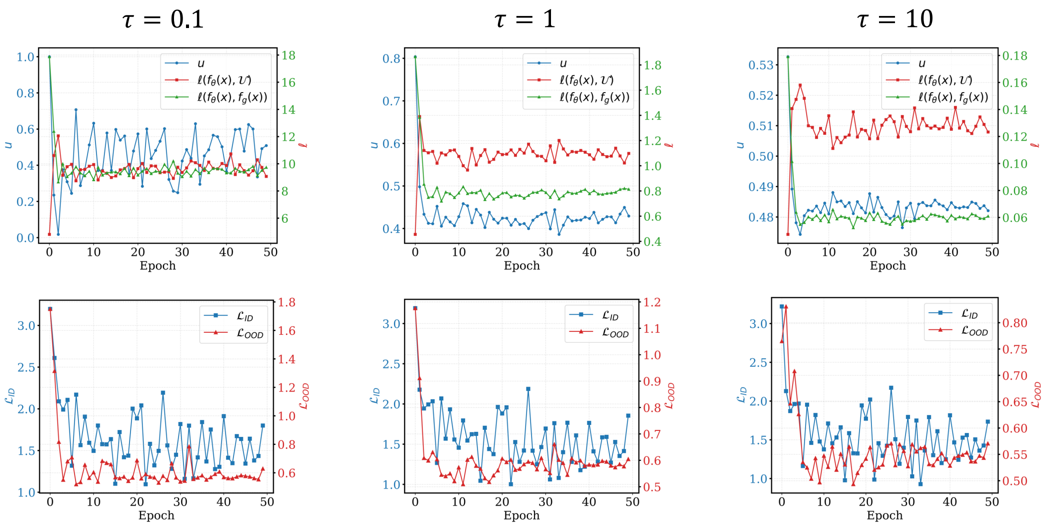

Figure 1: More hyperparameter τ visualization results.

APPENDIX H THE EXPERIMENT ON GEB COMPONENT CONSTANT.

This part shows the results of C on the test set. Obviously, for LoCoOP, the KL divergence for general knowledge is large. For our method, both KL divergences on the test set remain small, which explains one of the reasons why we consider these two terms as constant terms in the generalization error

Table 8: Two KL divergence results on the test set.

Method	$KL(f_\theta(x),\mathcal{U})$	$KL(f_\theta(x), f_g(x))$
LoCoOp	0.84	1.89
Ours	0.81	1.13

APPENDIX I OOD DATASETS.

iNaturalist. The dataset under consideration is comprised of 859,000 biological specimens, which are divided into more than 5,000 taxonomic categories. The primary focus of the dataset is flora and fauna biodiversity. In accordance with the established protocol, the evaluation process is conducted using a sample of 10,000 images, selected at random from a total of 110 classes, with the exclusion of those that are already present in the ImageNet-1K database.

SUN. The scene recognition corpus under consideration contains 130,000 visual instances, which are divided into 397 environmental categories. For the purpose of comparative analysis, a curated subset of 10,000 images has been employed, sampled from 50 ImageNet-disjoint classes.

Places. Places provides complementary coverage of environmental semantics, mirroring SUN's conceptual scope in scene understanding. The assessment utilises 10,000 images from 50 non-overlapping classes.

TEXTURE. The present corpus is one that has been specifically compiled for the purpose of this study. It consists of 5,640 high-resolution texture patterns that have been organised into 47 material categories. A comprehensive evaluation is performed using the full dataset.

OpenImage-O. This rigorously curated visual recognition benchmark comprises 17,632 images that have been manually filtered through multi-stage quality assurance protocols, achieving 7.8× greater scale diversity than ImageNet-O through pixel-coverage optimisation.

SSB-hard. Derived from ImageNet-21K's hierarchical ontology through semantic scarcity sampling, this 49,000-image benchmark spans 980 visually complex categories characterised by high inter-class ambiguity.

NINCO. The dataset contains 5,879 meticulously annotated samples across 64 novel categories, thereby introducing conceptual novelty through systematic exclusion of ImageNet-1K semantic overlaps.

ImageNet-10. The creation of ImageNet-10 was driven by the necessity to emulate the class distribution of CIFAR-10, while incorporating high-resolution images. The following categories are contained within the dataset, along with their respective class identifiers: The following subject headings have been identified: The following terms are listed: 'warplane' (n04552348), 'sports car' (n04285008), 'brambling bird' (n01530575), 'Siamese cat' (n02123597), 'antelope' (n02422699). The following have been identified: 'Swiss mountain dog' (n02107574), 'bull frog' (n01641577), 'garbage truck' (n03417042), 'horse' (n02389026), and 'container ship' (n03095699).

ImageNet-20. In order to facilitate the evaluation of hard OODs with realistic datasets, ImageNet-20 has been curated. The dataset under consideration consists of 20 classes that are semantically similar to ImageNet-10. The categories are selected based on the distance in the WordNet synsets. The following categories are contained therein: The following items are listed herewith: The following objects are documented: a sailboat (n04147183), a canoe (n02951358), a balloon (n02782093), a tank (n04389033), a missile (n03773504), and a bullet train (n02917067). The following species were documented: A starfish (n02317335), a spotted salamander (n01632458), a common newt (n01630670), a zebra (n01631663), and a frilled lizard (n02391049). For the purposes of this study, the following taxa were selected: the green lizard (n01693334), the African crocodile (n01697457), the Arctic fox (n02120079), the timber wolf (n02114367), the brown bear (n02132136), the moped (n03785016), the steam locomotive (n04310018), the space shuttle (n04266014) and the snowmobile (n04252077).

REFERENCES

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition*, Oct 2020.

Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. *Advances in neural information processing systems*, 35:35087–35102, 2022.

Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Gl-mcm: Global and local maximum concept matching for zero-shot out-of-distribution detection. *International Journal of Computer Vision*, pp. 1–11, 2025.

Shuhuai Ren, Aston Zhang, Yi Zhu, Shuai Zhang, Shuai Zheng, Mu Li, Alexander J Smola, and Xu Sun. Prompt pre-training with twenty-thousand classes for open-vocabulary visual recognition. *Advances in Neural Information Processing Systems*, 36:12569–12588, 2023.

Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3903–3911, 2020.