A Qualitative Analysis

A.1 Case study

In Figure 1, we visualize the top-1 retrieved images for given text queries in 11 languages on XTD dataset [1]. Compared with the multilingual vision-language pre-training model UC^2 [17], MLA can better capture entities, attributes, and actions to retrieve the correct image. Specifically, given simple queries that contain few entities such as Query #1 or Query #2, the images retrieved by MLA show high consistency across languages, since the representations of non-English queries are aligned to English in the NLT stage. For the more complex queries such as Query #3 or Query #4, MLA also shows better fidelity to all entities in most cases.

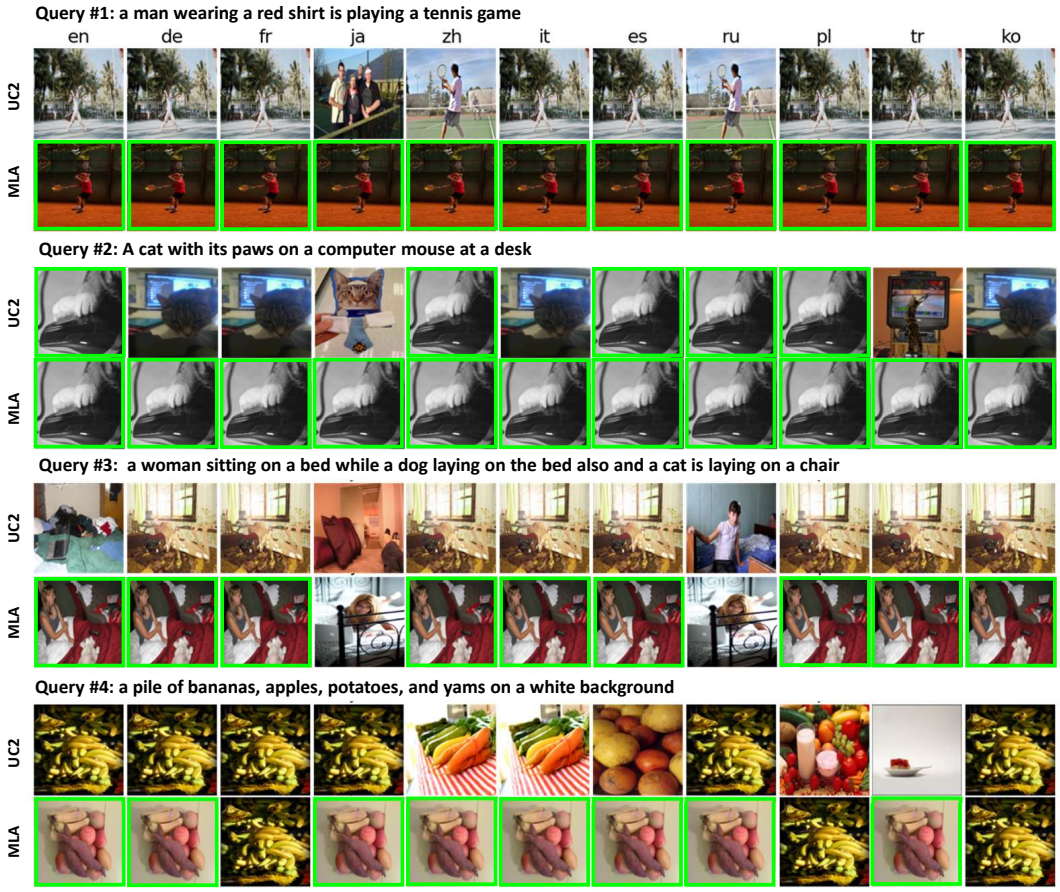

Figure 1: Top-1 retrieved images for given text queries in 11 languages on XTD dataset. Only English queries are shown in this figure. The correct images are bordered green.

A.2 Representation visualization

To visualize the multimodal and multilingual representation space, we translate the English class labels of CIFAR10 [10] into 5 languages including German (de), French (fr), Czech (cs), Chinese (zh), and Japanese (ja). The images and labels in 6 languages are encoded into representations through MLA_{CLIP}. Figure 2 shows the t-SNE [16] visualization of these representations. We can see that the representations from different languages and modalities are clustered according to the semantics. It suggests that MLA_{CLIP} indeed can project images and multilingual sentences into a shared multimodal and multilingual space.

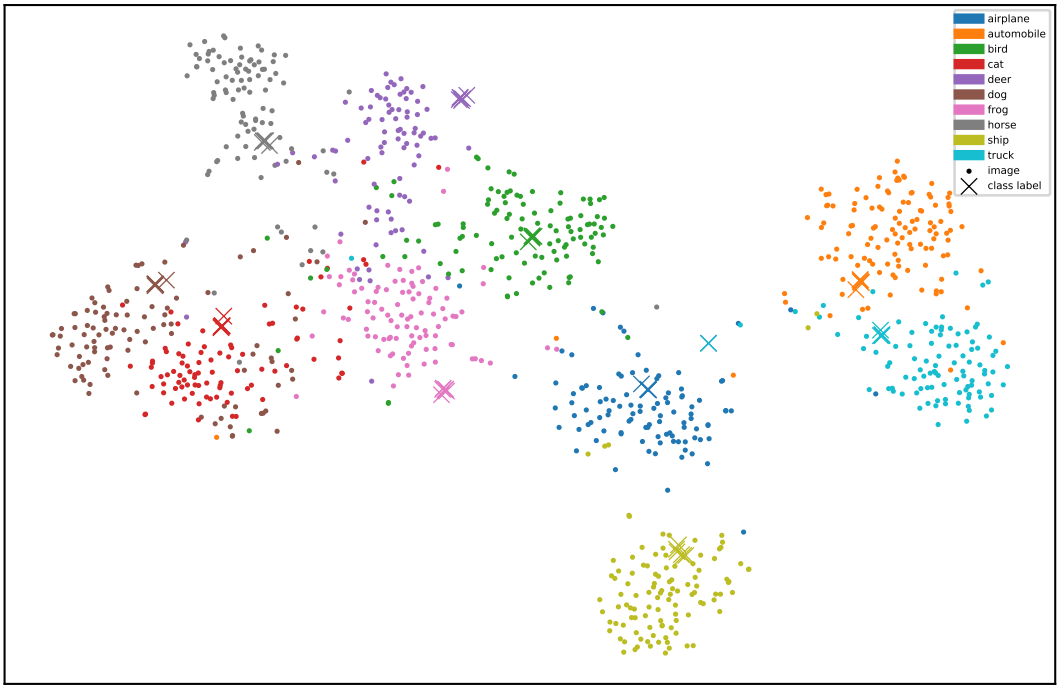

Figure 2: Representation visualization with t-SNE. The categories are color coded. '•' denotes a image representation, and '×' denotes a class label representation in a certain language.

B Additional Ablation studies

We conduct additional ablation studies to verify the effectiveness of MLA. All experiments in this section are conducted on zero-shot image-text retrieval.

B.1 Structure of language acquirer

In our proposed MLA, we implement the language acquirer as a bottleneck MLP. In Table.1, we compare the different structure of the language acquirer, the bottleneck MLP and a linear projection layer with the same amount of parameters. MLP works slightly better than the linear projection. Thus, we choose MLP to conduct our major experiments.

Table 1: Ablation study on structure of language acquirer.

Method	Component	Multi30K			MSCOCO 1K	
		de	fr	cs	ja	zh
MLA_{CLIP}	Linear	78.2	77.6	69.3	74.6	78.0
MLA_{CLIP}	MLP	**78.7**	**77.7**	**70.8**	**74.9**	**78.5**

B.2 MLA on CLIP with different structures

We additionally apply MLA to CLIP [11] in different sizes with two kinds of structures: ResNet [6] and ViT [4]. The results in Table 2 indicate that MLA can perform better on all languages when stronger monolingual VLPs are provided.

Table 2: Applying MLA on CLIP with different structures and sizes.

Structure	Multi30K				MSCOCO 1K		
	en	de	fr	cs	en	ja	zh
ResNet50	84.2	76.6	75.8	67.5	78.3	72.7	75.9
ResNet101	83.9	76.9	77.3	70.4	78.9	73.1	76.9
ResNet50x4	86.0	80.7	80.3	73.1	80.4	75.5	78.2
ResNet50x16	87.8	80.6	79.9	73.8	81.7	74.4	77.6
ResNet50x64	**89.9**	**84.2**	**84.1**	**78.1**	**82.2**	**79.3**	**80.6**
ViT-B-32	84.4	78.7	77.7	70.8	79.4	74.9	78.5
ViT-B-16	86.4	80.8	80.9	72.9	80.9	76.7	79.2
ViT-L-14	**87.9**	**83.1**	**83.5**	**77.0**	**82.5**	**78.5**	**79.1**

B.3 Objectives in the two-stage training

In the default setting, we use the MSE objective during the NLT stage and the NCE objective [5] during the LE stage. The MSE objective requires paired representations to be completely consistent, while the NCE objective only requires positive pairs to be closer than negative ones. We conduct experiments to use different objectives in the two stages. As shown in Table 3, we observe that the MSE objective is more suitable for the NLT (row 1 vs. row 2, row 7 vs. row 8) stage, and the NCE objective performs better for the LE stage (row 3 vs. row 4, row 5 vs. row 6). We consider the reason is that in the NLT stage, we leverage translation pairs to build alignment between languages. Since the two sentences of a translation pair are highly semantically related, their representations can be very similar. Thus, optimizing a strong objective like MSE during the NLT stage is feasible. However, during the LE stage, the optimization is conducted with image-text pairs. Although the image and text are semantically related, one sentence can hardly describe all the information in the image. Therefore, a weak objective like NCE is suitable for the LE stage.

Table 3: Ablation study on objectives in the two training stages. mse: MSE objective, nce: NCE objective

Row	Stage one		Stage two		Multi30K			MSCOCO 1K	
	NLT	LE	NLT	LE	de	fr	cs	ja	zh
1	mse				76.3	74.2	67.2	72.1	75.7
2	nce				63.0	58.5	49.6	57.6	64.8
3		mse			47.2	47.0	37.4	46.3	54.9
4		nce			68.2	67.7	58.6	65.9	71.7
5	mse			mse	55.0	51.3	43.8	50.9	57.9
6	mse			nce	**78.7**	**77.7**	**70.8**	**74.9**	**78.5**
7	mse		mse	nce	78.4	77.3	69.9	74.2	78.1
8	mse		nce	nce	78.1	77.2	69.5	73.9	78.2

B.4 Multilingual Acquisition vs. Cross-modal Acquisition

MLA adopts the "Multimodal→Multilingual" strategy that empowers VLP models with multilingual capability. However, there is another option of "Multilingual→Multimodal" that empowers multilingual pre-training models with multimodal capability. To make a comparison between these two strategies, we implement the Cross-Modal Acquisition (CMA) that inserts cross-modal acquirers in each layer of the multilingual pre-training model M-BERT [3]. We keep the pre-trained M-BERT

fixed and train the cross-modal acquirers with the same two-stage strategy as MLA. From Table 4, we find that CMA performs worse than MLA in all languages. It suggests that generalizing multilingual models to multimodal is harder than generalizing multimodal models to multilingual through lightweight acquirers.

Table 4: Multilingual Acquisition vs. Cross-modal Acquisition

Method	Multi30K				MSCOCO 1K		
	en	de	fr	cs	en	ja	zh
CMA_{CLIP}	80.2	73.9	72.8	67.0	76.3	69.8	75.1
MLA_{CLIP}	**84.4**	**78.7**	**77.7**	**70.8**	**79.4**	**74.9**	**78.5**

B.5 Details of implementing MURAL

We implement MURAL [9] on the 6 languages considering our computing budgets. The dual-encoders of MURAL are implemented with CLIP-ViT-32 and M-BERT-base [3] respectively, since we find that initializing the dual-encoders with both pre-trained models can boost the performance. We train MURAL on CC300K (same as MLA) using 1 V100 GPU with a batch size of 128, and on CC3M [14] (the largest dataset we can access) using 8 V100 GPUs with a batch size of 1024. The learning rate is set to 1e-4. Both models converge in about 1 day and 4 days respectively. The results are shown in Table 5. It indicates that even initializing the dual-encoders, MURAL performs worse than MLA. Note that under the fair comparison, MLA also shows its low-cost merit, since the data and computing resources of MURAL pre-trained on CC3M are much larger.

Table 5: Comparing with MURAL pre-trained with different data and initialization.

Method	Data	Initializing		Multi30K				MSCOCO 1K		
		M-BERT	CLIP	en	de	fr	cs	en	ja	zh
MURAL	CC300K	✗	✗	23.0	20.8	19.6	17.5	29.9	26.3	31.7
MURAL	CC300K	✗	✓	59.5	55.8	52.6	47.2	63.5	56.8	75.1
MURAL	CC300K	✓	✓	67.8	62.7	60.8	57.5	68.1	62.5	67.0
MURAL	CC3M	✓	✓	79.3	73.7	72.4	69.2	76.1	71.1	74.9
MLA_{CLIP}	CC300K	✓	✓	**84.4**	**78.7**	**77.7**	**70.8**	**79.4**	**74.9**	**78.5**

B.6 Experiment on WIT

We conduct an evaluation on WIT [15] dataset to further examine MLA on real low-resource languages. WIT [15] contains Wikipedia-based image-text pairs in 108 languages. We follow the test set proposed in the IGLUE benchmark [2] that contains 500-1000 image-text pairs in 10 languages. We train both MLA and MURAL(pre-trained on CC300K and CC3M) in the 10 languages with CC69L and perform the evaluation on WIT directly. As shown in Table 6, MLA still outperforms MURAL in most languages on this benchmark, which validates the effectiveness of MLA.

Table 6: Evaluation on the WIT dataset.

Method	Data	ar	bg	da	el	et	id	ja	ko	tr	vi	mean
MURAL	CC300K	26.2	22.9	26.8	28.3	12.6	25.0	16.1	18.6	25.8	30.5	23.3
MURAL	CC3M	27.9	25.1	28.4	**30.1**	13.6	**27.1**	16.6	**20.3**	**28.8**	32.2	25.0
MLA_{CLIP}	CC300K	**30.7**	**25.3**	**30.8**	29.9	**14.3**	26.7	**17.0**	19.8	28.1	**34.3**	**25.7**

C Open-domain Image Classification

In order to test the open-domain capability of models, we conduct zero-shot open-domain image classification experiments on CIFAR100 [10], ImageNet-V2 [12], ImageNet-R [7] and ImageNet-

A [8] datasets. As shown in Table 7, MKD [13] performs badly on open-domain image classification. We consider the reason is that MKD abandons the original text encoder which contains open-domain multimodal knowledge from large-scale pre-training. In contrast, MLA keeps the original text encoder fixed and thus could maintain the open-domain capability of the pre-training model.

Table 7: Top-1 Accuracy of zero-shot open-domain image classification.

Method	CIFAR100	ImageNet-V2	ImageNet-R	ImageNet-A
MKD_{CLIP}	32.8	54.7	37.7	23.5
MLA_{CLIP}	**64.2**	**63.4**	**69.0**	**31.4**

References

[1] Pranav Aggarwal, Ritiz Tambi, and Ajinkya Kale. Towards zero-shot cross-lingual image retrieval and tagging. *CoRR*, abs/2109.07622, 2021.

[2] Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. IGLUE: A benchmark for transfer learning across modalities, tasks, and languages. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2370–2392. PMLR, 17–23 Jul 2022.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, 2019.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[5] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition*, 2016.

[7] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *IEEE/CVF International Conference on Computer Vision*, 2021.

[8] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[9] Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. MURAL: Multimodal, multitask representations across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3449–3463, 2021.

[10] A Krizhevsky. Learning multiple layers of features from tiny images. *Master's thesis, University of Tront*, 2009.

[11] 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. Learning transferable visual models from natural language supervision. In *38th International Conference on Machine Learning*, volume 139, pages 8748–8763, 2021.

[12] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*. PMLR, 2019.

[13] Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, 2020.

[14] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[15] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2449, 2021.

[16] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[17] Mingyang Zhou, Luowei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. Uc2: Universal cross-lingual cross-modal vision-and-language pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4155–4165, June 2021.