Supplementary Materials of "Self-Adapting Large Visual-Language Models to Edge Devices across Visual Modalities"

Kaiwen Cai[1], Zhekai Duan[1], Gaowen Liu[2], Charles Fleming[2], and Chris Xiaoxuan Lu[3]

[1] University of Edinburgh, UK
[2] Cisco Research, USA
[3] University College London, UK

In the supplemental material included, we offer further information regarding the quantization procedure and the approach to quantization-aware training implemented for the student model. Additionally, we outline the methodology utilized for creating label supersets with the assistance of ChatGPT-4 [1]. We also include experimental results concerning the overall efficacy of the student model, alongside an ablation study. To enhance comprehension of the model's effectiveness across various visual modalities, we conclude with visual representations of the open vocabulary classification outcomes on multiple datasets.

A Model Quantisation

We consider integer uniform quantization [11] which maps high-precision model weights and activations to low-precision representations. The quantization process is presented as follows:

$$\begin{aligned} s &= \frac{2^{b-1}-1}{\alpha} \\ \texttt{Quantization}: x_q &= \texttt{clip}(\lceil x \cdot s \rfloor, -2^{b-1}+1, 2^{b-1}-1) \\ \texttt{De-quantization}: \widetilde{x} &= \frac{x_q}{s} \end{aligned} \tag{1}$$

where α is the maximum representable value, s is the scale factor, b is the bit-width, x is the original weight or activation, x_q is the quantized weight or activation, $\widetilde{x}$ is the weight or activation restored from quantized ones, and `clip` is the clipping function.

$$\texttt{clip}(x, a, b) = \begin{cases} a & \text{if } x < a \\ b & \text{if } x > b \\ x & \text{otherwise} \end{cases} \tag{2}$$

Per-tensor quantization applies the same scale factor across the entire tensor, whereas per-channel quantization assigns a distinct scale factor to each channel. The scale factor is determined using a calibration set in a static quantization

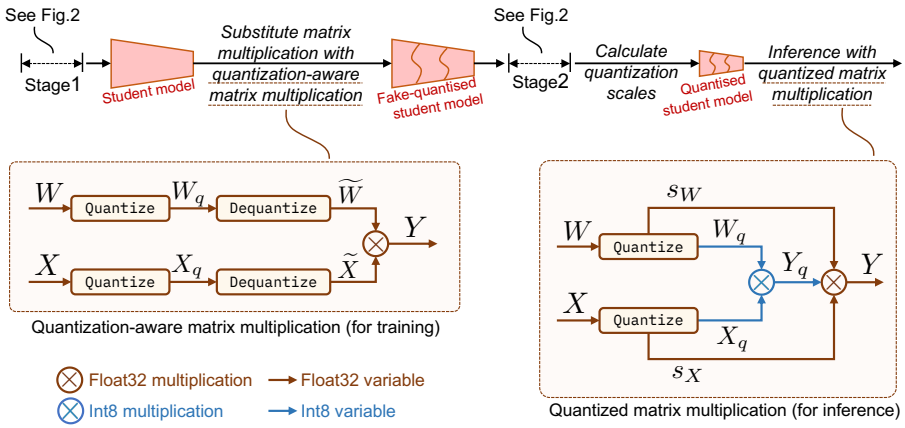

Fig. A: Quantization-aware matrix multiplication is used to take-quantize the student model

approach or computed on the fly during inference for dynamic quantization. Quantizing the weights and activations of a Float32 model to Int8 reduces both the model size and inference latency. We follow the common practice [10] [11] and apply per-channel quantization to the weights and per-tensor quantization to the activations. When evaluating static quantization performance, we follow [11] and calculate the scale factors using a subset of the training set, *e.g.*, 64 pairs of RGB and non-RGB images.

Figure A depicts the computation flow of quantization-aware matrix multiplication during training and quantized matrix multiplication during inference. After training the student model in stage-1, the fake-quantized student model is obtained by including the `Quantize` and `Dequantize` operations in the model. After stage-2, quantization scales are computed using a subset of the training data. For deployment on edge devices, the fake-quantized student model is transformed into the quantized student model using TensorRT, replacing quantization-aware matrix multiplication with quantized matrix multiplication.

B Label Supersets Generation

Scene classification typically involves categorizing scenes within the domain of indoor, outdoor, or remote sensing. Acknowledging this, we create a comprehensive set of labels tailored to various scene classification challenges. For instance, we might develop a set of indoor scene labels for the ScanNet dataset and a distinct set of labels for satellite imagery in the EuroSAT dataset. In particular, we utilize ChatGPT-4 [1] to expertly craft these extensive label supersets.

For the ScanNet dataset, we use the following prompt:

```
User: Please generate a list of indoor scene categories without
descriptions.
------------------------------------------------
ChatGPT-4: Living Room, Kitchen, Bedroom, Bathroom, Dining Room,
Home Office, Basement, Attic, Office and Cubicles, Conference Room,
Classroom, Laboratory, Library, Gym/Fitness Center, Shopping Mall,
Retail Store, Hotel Lobby, Hotel Room, Restaurant, Cafe, Bar or
Nightclub, Cinema/Theater, Airport Terminal, Subway Station, Train
Station, Bus Interior, Hospital Ward, Operating Room, Dental Clinic,
Veterinary Clinic, Museum, Art Gallery, Church, Temple, Mosque,
Classroom, Lecture Hall, Laboratory, School Hallway, Dormitory,
Indoor Pool, Bowling Alley, Arcade, Indoor Playground, Courthouse,
City Hall, Police Station, Fire Station, Factory Floor, Warehouse,
Server Room.
```

For the EuroSAT dataset, we use the following prompt:

```
User: Please generate a list of satellite scene categories without
descriptions.
------------------------------------------------
ChatGPT-4: Urban or Built-up Areas, Agriculture, Forest, Grassland,
Water Bodies, Wetlands, Desert, Snow and Ice, Bare Land or Barren,
Mountainous Terrain, Coastal Zones, Riparian Zones, Glacial and
Periglacial Landscapes, Volcanic Features, Urban Green Spaces.
```

C Results

C.1 Overall Accuracy

To provide a comprehensive performance evaluation of `EdgeVL` compared to existing methods, we present a detailed accuracy analysis in Tab. A. Except for a small performance difference to CQD [8] using Swin-T on EuroSAT (61.3% vs. 62.5%), `EdgeVL` demonstrates superior performance across all backbones and on two datasets, surpassing all other methods by a significant margin.

C.2 Quantization Aware Contrastive Learning

We compare the performance of `EdgeVL` with other quantization methods when dynamic quantization is applied during inference. Comparison between Tab. 3 in the main script and Tab. B reveals that dynamic quantization outperforms static quantization. This finding is also supported by [11]. Just like the results showcased in Tab. 3 in the main script, `EdgeVL` emerges as the most effective method in maintaining model performance post-quantization.

Table A: Overall accuracy comparison. and denote the top1 accuracy of non-RGB and RGB images, respectively. And denotes the average of the two, all in percentage. The same applies to the following tables. Best viewed in color.

	Methods	Bits	ScanNet (%) ↑			EuroSAT (%) ↑		
	Pretrained CLIP-B [7]	F32	4.5	36.2	20.4	16.8	40.4	28.6
	Pretrained CLIP-G [7]	F32	6.2	47.3	26.8	16.9	54.0	35.5
DAT-T	Frank [2]	F32	8.3	21.7	15.0	32.4	31.4	31.9
	Gupta [3]	F32	16.8	22.4	19.6	25.4	27.2	26.3
	CMKD [4] (non-RGB)	F32	37.8	11.5	24.6	36.7	18.9	27.8
	CMKD [4] (RGB)	F32	4.0	42.5	23.2	15.2	40.9	28.1
	Fida [9]	F32	38.9	5.8	22.3	53.7	14.1	33.9
	CQD [8]	F32	40.1	6.7	23.4	37.2	17.5	27.3
	SKD [12]	F32	21.2	40.7	31.0	24.8	41.5	33.2
	`EdgeVL`	Int8	**47.9**	**52.0**	**49.9**	**61.0**	**65.7**	**63.3**
Swin-T	Frank [2]	F32	8.0	14.8	11.4	41.9	39.0	40.5
	Gupta [3]	F32	22.0	17.5	19.8	45.8	39.5	42.6
	CMKD [4] (non-RGB)	F32	38.9	4.0	21.4	59.5	22.0	40.8
	CMKD [4] (RGB)	F32	3.2	42.4	22.8	20.1	62.4	41.2
	Fida [9]	F32	41.2	1.3	21.3	54.5	10.5	32.5
	CQD [8]	F32	41.4	4.9	23.1	**62.5**	18.3	40.4
	SKD [12]	F32	31.2	37.8	34.5	24.1	28.8	26.4
	`EdgeVL`	Int8	**46.0**	**48.7**	**47.4**	61.3	**67.1**	**64.2**
ViT-S	Frank [2]	F32	11.3	14.0	12.7	49.2	37.9	43.5
	Gupta [3]	F32	19.6	14.9	17.2	54.2	42.4	48.3
	CMKD [4] (non-RGB)	F32	38.0	4.7	21.3	61.2	34.4	47.8
	CMKD [4] (RGB)	F32	3.0	42.5	22.8	14.3	60.4	37.3
	Fida [9]	F32	40.0	4.5	22.2	56.7	20.3	38.5
	CQD [8]	F32	38.0	4.0	21.0	62.4	36.4	49.4
	SKD [12]	F32	28.7	37.6	33.1	22.9	50.3	36.6
	`EdgeVL`	Int8	**42.0**	**47.5**	**44.7**	**62.9**	**66.8**	**64.8**

Table B: QAT meets Contrastive Learning. + denotes in combination with stage-1. Inferencing with Dynamic quantization.

Methods	Bits	DAT-T (%)			Swin-T (%)			ViT-S (%)		
Stage-1	F32	38.6	40.6	39.6	39.9	41.2	40.5	37.8	40.7	39.3
+PTQ [5]	Int8	35.1	39.1	37.1	32.0	35.0	33.5	26.6	28.5	27.6
+QAT [5]	Int8	39.6	41.3	40.5	39.6	39.5	39.5	38.4	41.9	40.1
+QViT [6]	Int8	38.8	41.1	39.9	37.8	39.9	38.8	31.9	35.2	33.5
+Stage-2	Int8	**49.0**	**51.5**	**50.2**	**47.6**	**49.7**	**48.7**	**44.9**	**49.1**	**47.0**

C.3 Effect of Finetuning CLIP

To further demonstrate the effectiveness of our fine-tuning method, we use the CLIP-B model as the student model and the CLIP-G model as the teacher model. The CLIP-B model was trained following the same approach as in stage-1 of Fig. 1, with training parameters identical to those in Sec. 4.1. As shown in Tab. C, the fine-tuned CLIP-B model exhibits improved performance compared to the pre-trained models on non-RGB modalities, but its performance still lags much behind `EdgeVL`. We attribute this to the fact that the CLIP-B model is designed for large-scale image-text retrieval tasks, meaning it excels with larger datasets. In contrast, `EdgeVL` is specifically optimized for cross-modal classification tasks. Additionally, the contrastive learning process during stage 2 helps `EdgeVL` to learn better feature representations.

Table C: Effect of finetuning CLIP-B.

Methods	Bits	ScanNet (%) ↑			EuroSAT (%) ↑		
Finetuned CLIP-B	F32	27.1	27.9	27.5	46.9	47.7	47.3
`EdgeVL` (DAT-T)	Int8	**47.9**	**52.0**	**49.9**	**61.0**	**65.7**	**63.3**

C.4 Qualitative Analysis

Fig. B: Visualization of the predictions of different models on ScanNet. Red and green colors indicate incorrect and correct predictions, respectively.

Figure B and Figure C illustrate visual examples of the predictions made by CLIP-G [7], CQD [8], SKD [12], and `EdgeVL` (ViT-S) on the ScanNet and EuroSAT datasets, respectively. The visualizations underscore the superior accuracy of `EdgeVL` over competing methods, particularly with non-RGB images.

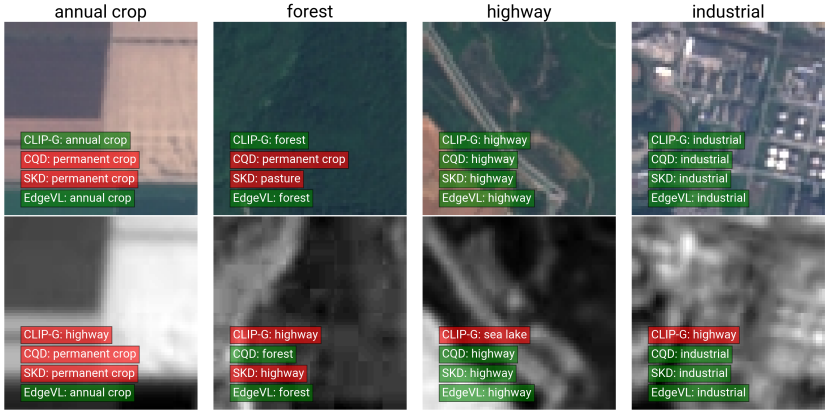

Fig. C: Visualization of the predictions of different models on EuroSAT. Red and green colors indicate incorrect and correct predictions, respectively.

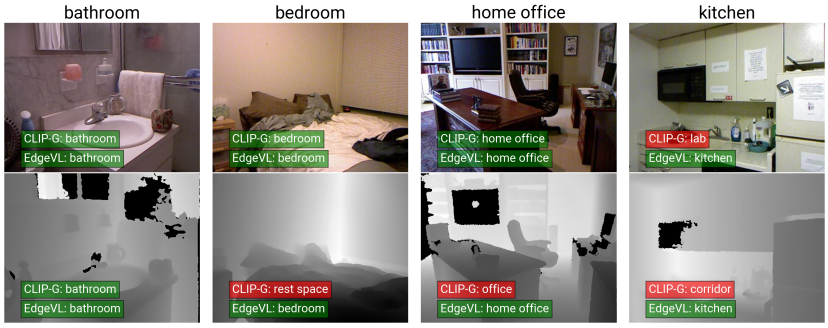

Fig. D: Visualization of the predictions of different models on SUNRGBD. Red and green colors indicate incorrect and correct predictions, respectively.

Fig. E: Visualization of the predictions of different models on SUNRGBD. Red and green colors indicate incorrect and correct predictions, respectively.

Figure D and Figure E depict the cross-dataset predictions on SUNRGBD by `EdgeVL` (DAT-T) that was trained on ScanNet. In Figure D, `EdgeVL` demonstrates good performance in distinctive scenes like bathrooms and kitchens. However, in settings such as furniture stores and lecture theatres shown in Figure E, `EdgeVL` struggles even with the RGB modality, while CLIP-G achieves higher accuracy in these instances. This is due to the infrequent presence of these scenarios (e.g., furniture stores and lecture theatres) in the ScanNet dataset, making them unfamiliar to the model. We also discussed the potential reason for data amount in our main paper. Our future research will focus on improving the model's generalization ability with limited adapataion data.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., others: Gpt-4 technical report. arXiv (2023)
2. Hafner, F.M., Bhuyian, A., Kooij, J.F., Granger, E.: Cross-modal distillation for RGB-depth person re-identification. Computer Vision and Image Understanding **216**, 103352 (2022)
3. Hoffman, J., Gupta, S., Leong, J., Guadarrama, S., Darrell, T.: Cross-modal adaptation for RGB-D detection. In: 2016 IEEE International Conference on Robotics and Automation (ICRA). pp. 5032–5039. IEEE (2016)
4. Hong, Y., Dai, H., Ding, Y.: Cross-Modality Knowledge Distillation Network for Monocular 3d Object Detection. In: European Conference on Computer Vision (ECCV) (2022)
5. Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., Kalenichenko, D.: Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2704–2713. IEEE (2018)
6. Li, Y., Xu, S., Zhang, B., Cao, X., Gao, P., Guo, G.: Q-ViT: Accurate and Fully Quantized Low-bit Vision Transformer. In: Conference on Neural Information Processing Systems (NeurIPS). vol. 35, pp. 34451–34463 (2022)
7. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: [CLIP] Learning Transferable Visual Models From Natural Language Supervision. In: International Conference on Machine Learning (ICML). pp. 8748–8763 (2021)
8. Su, J.C., Maji, S.: Adapting Models to Signal Degradation using Distillation. In: British Machine Vision Conference (BMVC) (2017)
9. Thoker, F.M., Gall, J.: Cross-Modal Knowledge Distillation for Action Recognition. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 6–10. IEEE (2019)
10. Wu, H., Judd, P., Zhang, X., Isaev, M., Micikevicius, P.: Integer Quantization for Deep Learning Inference: Principles and Empirical Evaluation. arXiv **abs/2004.09602** (2020)
11. Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., Han, S.: Smoothquant: Accurate and Efficient Post-Training Quantization for Large Language Models. In: International Conference on Machine Learning (ICML). pp. 38087–38099 (2023)
12. Yang, C., An, Z., Zhou, H., cai, l., Zhi, X., Wu, J., xu, y., Zhang, Q.: Mixskd: Self-Knowledge Distillation from Mixup for Image Recognition. In: European Conference on Computer Vision (ECCV) (2022)