Boost Embodied AI Models with Robust Compression Boundary

Chong Yu[1] , **Tao Chen**[2,*] , **Zhongxue Gan**[1,*]
[1]Academy for Engineering and Technology, Fudan University
[2]School for Information Science and Technology, Fudan University
21110860050@m.fudan.edu.cn, {eetchen, ganzhongxue}@fudan.edu.cn

Abstract

The rapid improvement of deep learning models with the integration of the physical world has dramatically improved embodied AI capabilities. Meanwhile, the powerful embodied AI models and their scales place an increasing burden on deployment efficiency. The efficiency issue is more apparent on embodied AI platforms than on data centers because they have more limited computational resources and memory bandwidth. Meanwhile, most embodied AI scenarios, like autonomous driving and robotics, are more sensitive to fast responses. Theoretically, the traditional model compression techniques can help embodied AI models with more efficient computation, lower memory and energy consumption, and reduced latency. Because the embodied AI models are expected to interact with the physical world, the corresponding compressed models are also expected to resist natural corruption caused by real-world events such as blur, darkness, weather conditions, and even adversarial corruption. This paper explores the novel paradigm to boost the efficiency of the embodied AI models and the robust compression boundary. The efficacy of our method has been proven to find the optimal balance between accuracy, efficiency, and robustness in real-world conditions.

1 Introduction

As of 2025, the field of embodied AI is witnessing significant advancements, driven by innovations in technology and a growing understanding of how to integrate AI with physical systems. For example, various classification [Yang *et al.*, 2023b], object detection [Jiang *et al.*, 2023], and semantic segmentation [Liu *et al.*, 2023b] models help autonomous vehicles perceive the surrounding environments better. Visual SLAM and point cloud models further improve autonomous vehicles' positioning and navigation capabilities in a dynamic environment.

Another good example is introducing the vision and language foundation models [Touvron *et al.*, 2023] into the control strategies learning for robotic manipulation. The enormous robot manipulation datasets [Khazatsky *et al.*, 2024] only have 100K to 1M examples, while the foundation models are trained with Internet-scale pretraining datasets [Laurençon *et al.*, 2022]. By fully utilizing the generalization capabilities of the foundation models, the robots may be able to extrapolate behaviors to different surrounding conditions (e.g., manipulation positions [Brohan *et al.*, 2022], lighting [Chi *et al.*, 2023], scene distractors [Xie *et al.*, 2024]) and unseen situations (e.g., novel objects [Mees *et al.*, 2024], new task instruction [Walke *et al.*, 2023]). Following this paradigm, a series of the vision-language- action (VLA) models [Kim *et al.*, 2024] are explored to enhance the control and manipulation of robotics.

Figure 1: The autonomous driving models are expected to resist against the natural corruption caused by real-world events such as bad weather conditions and low-illumination in night.

The rapid development of large-scale neural models has broadly and profoundly influenced the capabilities of embodied AI. Meanwhile, large-scale neural models have drawbacks [Yu, 2021], such as high computational cost and energy consumption, limitations in response-sensitive scenarios, and high memory requirements in the edge platforms of embodied AI. Considering the application characters and actual costs, the computational core devices on embodied AI platforms tend to use embedded systems with low power consumption, limited computational power, and memory [Tinchev *et al.*, 2019] [Wan *et al.*, 2021]. So, model compression [Abbasi *et al.*, 2022] [Li *et al.*, 2022] is essential to alleviate the gaps between the increasing scale and considerable resource consumption of the models and their efficient deployment on the limited-capacity embodied AI hardware platforms.

Considering the safety and robustness concerns in the embodied AI application, e.g., autonomous driving area, when we want to compress an autonomous driving model into compressed form, we need to check further whether this com-

*Tao Chen and Zhongxue Gan are corresponding authors.

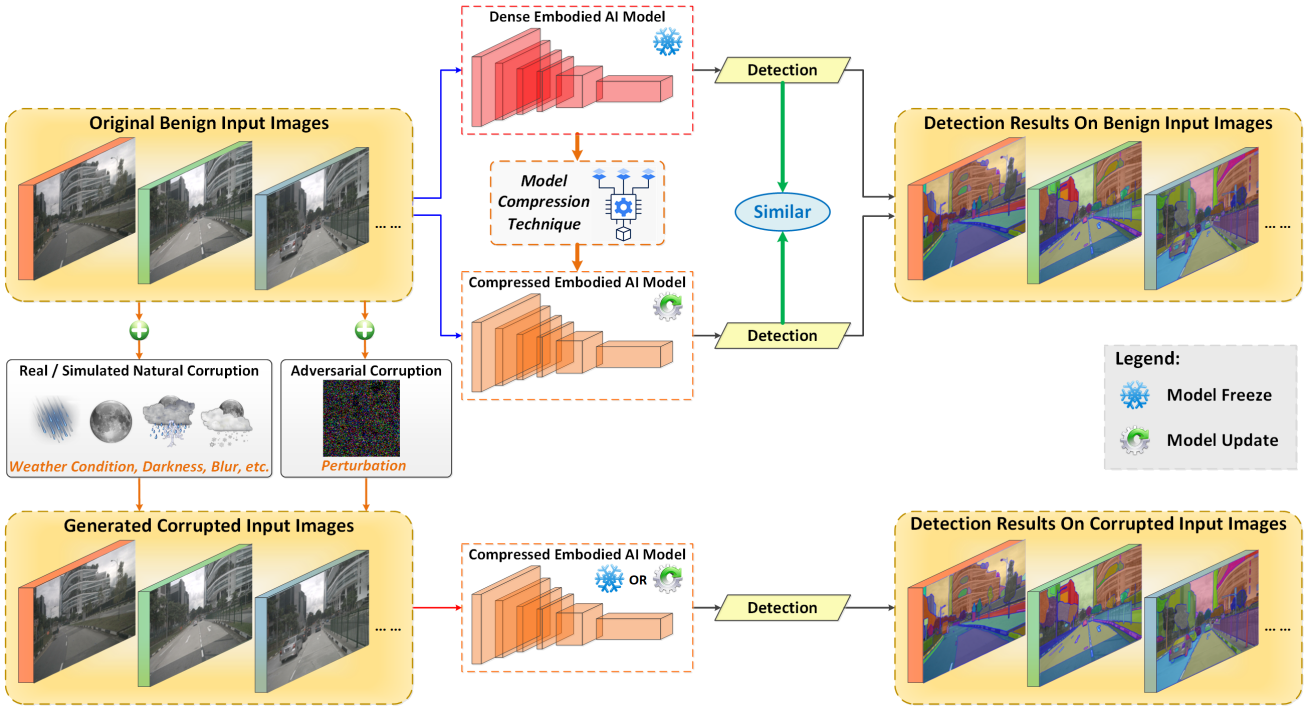

Figure 2: Naive model compression technique which disentangles the ***model compression process*** (***Upper***) on the benign training dataset with the ***against-corruption learning process*** (***Lower***) on the real or generated corrupted dataset.

pressed autonomous driving model is a qualified substitute facing corruptions. In traffic, the autonomous driving models are expected to resist natural corruption caused by real-world events shown in Figure 1 such as blur, darkness, weather conditions, and even the adversarial corruption to further guarantee the safety of the compressed embodied AI models.

In this paper, we first review the prior arts of model compression and their potential flaws when applied to embodied AI models, which require good accuracy, efficient deployment, safety, and robustness for corruption. Based on the analysis of past efforts, we explore and propose a novel paradigm for embodied AI models' compression. We push the robust compression boundary beyond the upper limitations that come from the original dense model. Finally, we evaluate our method on typical embodied AI models and benchmarking tasks to show the efficacy of our method for finding the optimal balance between accuracy, efficiency, and robustness in real-world conditions.

2 Related Work

2.1 Model Compression Against Corruption

When we try to use a compressed embodied AI model as the substitute for an original dense model, the typical strategy is applying the compression technique to ensure this obtained compressed model has a similar model accuracy and detection results with the original dense model on benign input images, shown in the **upper** part of the Figure 2. In this model compression process, only the weight parameters of the target compressed embodied AI model will be updated. The weight parameters of the original dense model are used in initialization or some knowledge-distillation-based compression techniques [Li *et al.*, 2023] [Yu *et al.*, 2023b] as the golden reference, so they are frozen.

After verifying the compressed models' quality on the benign dataset, the extra step is needed to evaluate whether the compressed is a qualified substitute when facing corruption, shown in the **lower** part of the Figure 2. Prior arts [Hnewa and Radha, 2020] will collect the real corrupted images or generate the simulated corrupted images to cover natural [Chen *et al.*, 2018] and adversarial corruption [Shen *et al.*, 2021] cases. Then, these corrupted input images are used to evaluate the compressed models' accuracy and detection results. In this case, the weight parameters of the target compressed embodied AI model will also be frozen. Some prior arts [Shen *et al.*, 2021] [Diffenderfer *et al.*, 2021] try to improve the robustness capabilities of the compressed embodied AI model, so they divide the subset of the collected corrupted images and unfrozen the weight parameters of the target compressed embodied AI model for against-corruption finetuning.

Some prior works further explore the relationship between model compression and robustness against corruption [Gui *et al.*, 2019] [Wang *et al.*, 2023]. Based on their exploration, the previous workflow shown in Figure 2 disentangles the model compression process with the against-corruption finetuning process while simplifying the complexity of the workflow but introducing flaws. The apparent flaw is that further finetuning on the corrupted samples may lead to the accuracy regression on the benign samples. Compensation efforts [Yan *et al.*, 2018] [Wang *et al.*, 2019] by recurrent training on individual benign and corrupted datasets or joint training on mixed

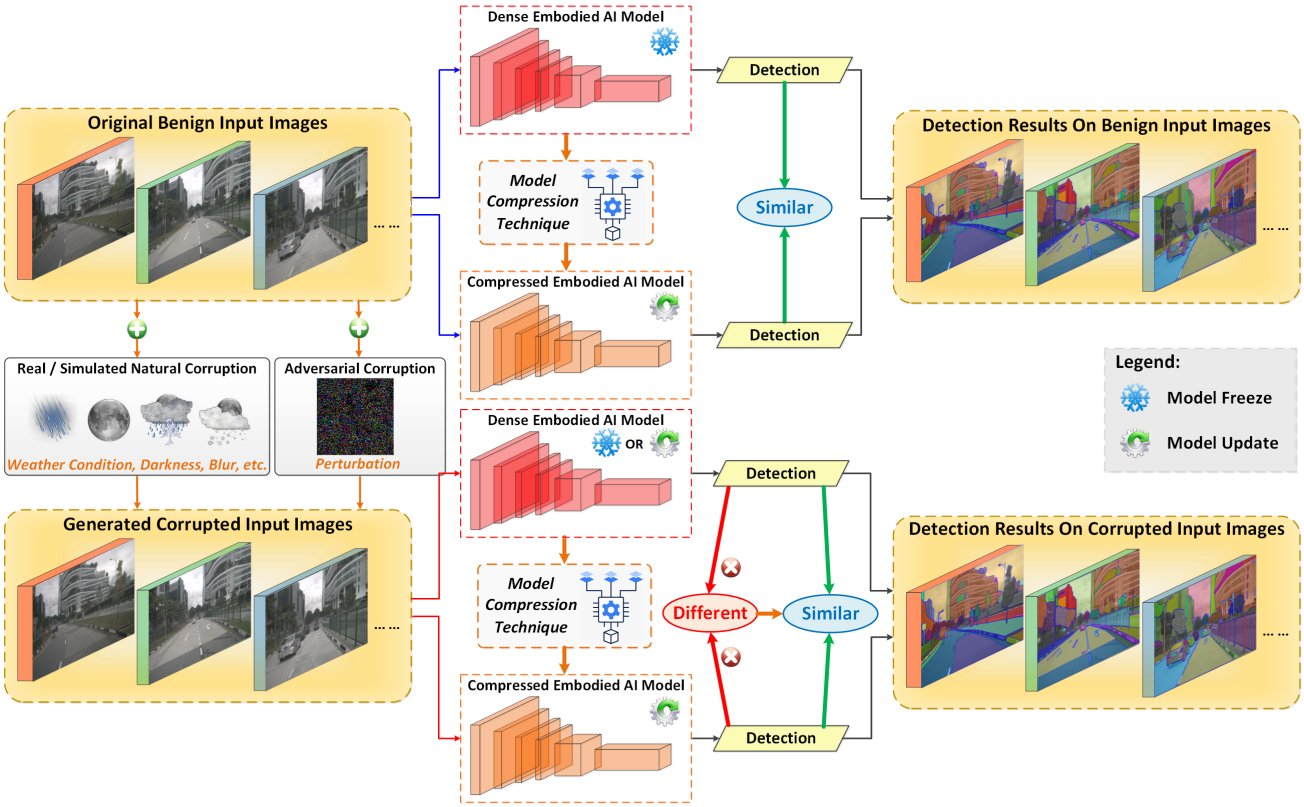

Figure 3: Efficient contrast-learning-based robust model compression and evaluation process to guarantee the compressed embodied AI model is a qualified substitute. (***Upper***) Check the detection results of the dense embodied AI model and the compressed counterpart, and guarantee they have the similar behaviors. (***Lower***) Collect the real or generate the simulated corrupted images to cover the natural and adversarial corruption cases. Check the detection results of the dense and the compressed counterpart on these corrupted images. Tuning the compressed model if they have different behaviors, and finally to guarantee they have the similar behaviors on the corrupted images as well.

benign and corrupted samples are proposed, but also raising concerns about more training cost and non-trivial efforts in balancing the accuracy and robustness. The more efficient workflow [Yu *et al.*, 2023a] introduces contrast learning into the robust model compression process, as shown in the Figure 3. During the normal model compression process, the workflow will check the detection results of the dense embodied AI model and the compressed counterpart on the corrupted images. If they have different behaviors, we need to tune the compressed model and finally guarantee they have similar behaviors on the corrupted samples, as well as on the benign samples.

2.2 Typical Embodied AI Models and Applications

Embodied AI represents a significant advancement in artificial intelligence, focusing on systems integrating cognitive abilities with physical actions to interact with and learn from their environments. Embodied AI has revolutionized autonomous vehicles by enabling them to navigate complex environments with human-like decision-making capabilities. Autonomous vehicles can adapt to changing road conditions and obstacles using advanced multi-modal perception systems. Robots equipped with embodied AI are used in manufacturing for assembly and material transport tasks. They can learn from their environment and adapt to changes in real time, improving efficiency. The system called Robotic Avatars [Luo *et al.*, 2022] allow users to operate robots remotely, capturing environmental data for telepresence applications. This technology has potential uses in telemedicine and hazardous environment operations.

With comparable and even superior effectiveness than the traditional convolution neural models, more transformer-based models [Liu *et al.*, 2023b] [Yang *et al.*, 2023a] [Liu *et al.*, 2023a] are explored and widely adopted in embodied AI typical applications with state-of-the-art performance. However, large-scale transformer-based models are computation-intensive and memory-intensive [Yu *et al.*, 2023b], placing an increasing burden on deployment efficiency.

3 Boost Robust Compression Boundary

By analyzing the prior model compression paradigms with a against-corruption setting, we find there are two main shortcomings in prior approaches, include:

- The model compression techniques do not fully dig into and utilize the difference between the benign and corrupted training samples.
- The compressed model cannot exceed the ***robustness boundary*** of the dense model because both the model

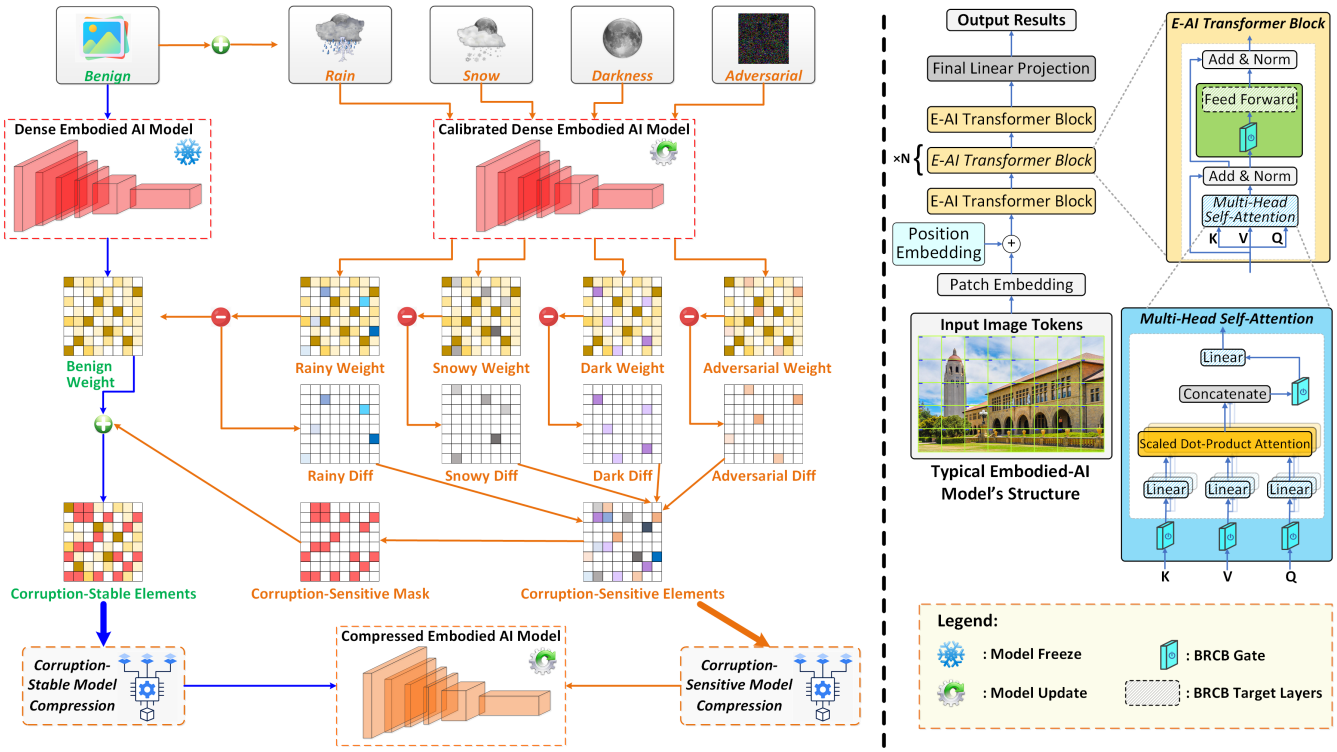

Figure 4: Workflow for **B**oost **R**obust **C**ompression **B**oundary (**BRCB**) **(Left)** *Against-Corruption Mechanism* to divide the original dense model into the corruptions-stable and corruptions-sensitive parts in fine-granularity. **(Right)** Push the limitation of *Robustness Boundary* of the dense model when facing corruption be introducing the ***BRCB gate layer*** in the *corruption-sensitive model compression* tuning process.

compression and contrast finetuning processes apply the dense model as the golden reference.

We intend to improve these two shortcomings and propose a new algorithm to compress the embodied AI models to push their robustness boundary and efficiency. The new algorithm is called **B**oost **R**obust **C**ompression **B**oundary (**BRCB**).

3.1 Against-Corruption Mechanism

Traditional studies regard corrupted samples as domain transformation. So, they usually use the corrupted samples for domain adaptation or mix them with benign ones in adversarial training. However, these prior solutions ignore whether such a difference between the corrupted and benign samples can be integrated and used for improving the model compression process. Inspired by this point, we design the against-corruption mechanism for **BRCB** workflow.

When we get the natural or adversarial corrupted samples, we first use them to calibrate the dense embodied AI model. Because the corrupted samples can be regarded as benign samples that concatenate the differences, the calibration, i.e., tiny tuning for the original dense embodied AI model, can usually work well with the corrupted samples. In other words, the differences between the original and the calibrated dense weight tensors should be marginal. So, we can disassemble the calibrated dense weight tensors as the similar weight part and the diff weight part, e.g., rainy diff, snowy diff, dark diff, and adversarial diff in the left side of Figure 4.

Then, the **BRCB** algorithm will concatenate all the diff weight tensors to generate a concentrated diff weight. All the elements in this concentrated diff weights are *corruption-sensitive*. With this corruptions-sensitive diff weight, we can transfer it as a corruptions-sensitive mask, using the bool elements to indicate the positions of the corruptions-sensitive elements. Because the corruptions-sensitive mask is in the same shape as the original dense model's weight, we can attach them to filter out a new weight tensor that contains all the *corruptions-stable* elements.

Because the input sample corruptions will less influence the elements in the corruptions-stable weight tensors, we can apply the combined sparse pruning and quantization techniques as normal for weight compression. And we call it the corruption-stable model compression path in the **BRCB** algorithm. Usually, the corruptions-stable weight tensors are more friendly to compression without out-of-range outliers. For the corruptions-sensitive diff weight, we will use the collected corrupted samples to execute the corruptions-sensitive model compression, with more details in the following section. Finally, the **BRCB** algorithm combines the compressed weights from corruptions-stable and corruptions-sensitive paths to obtain the final compressed embodied AI model. The detailed against-corruption mechanism is shown in the left side of Figure 4.

The key to **BRCB**'s against-corruption mechanism design is a fine-granularity division for the corruptions-stable and corruptions-sensitive elements from the original dense model. For the corruptions-stable elements, which are friendly for

Model	Method	Format	Benign Samples		Rainy Corruption		Snowy Corruption		Dark Corruption	
			NDS	mAP	NDS	mAP	NDS	mAP	NDS	mAP
BEVFormer	*Baseline*	FP32	51.7	41.6	41.4	32.2	42.5	32.9	37.3	27.5
	MiniViT	FP32	50.7	40.7	40.6	31.5	41.6	32.2	36.5	26.9
	UVC	FP32	49.8	39.7	39.9	30.7	40.9	31.4	35.9	26.3
	FQ-ViT	INT8	50.1	40.0	40.1	30.9	41.1	31.6	36.1	26.4
	PTQ-ViT	INT8	50.3	40.1	40.3	31.0	41.3	31.7	36.3	26.5
	GPUSQ-ViT	INT8	51.1	40.9	41.0	31.6	42.0	32.4	36.9	27.1
	BRCB	INT8	51.3	41.1	46.2	35.8	47.3	36.6	42.1	31.2
BEVFormer v2	*Baseline*	FP32	55.3	46.0	44.3	35.6	45.4	36.4	39.9	30.4
	MiniViT	FP32	54.2	45.0	43.4	34.8	44.5	35.6	39.1	29.7
	UVC	FP32	53.3	43.9	42.7	34.0	43.7	34.8	38.4	29.0
	FQ-ViT	INT8	53.6	44.2	42.9	34.2	44.0	35.0	38.6	29.2
	PTQ-ViT	INT8	53.8	44.3	43.1	34.3	44.2	35.1	38.8	29.3
	GPUSQ-ViT	INT8	54.7	45.3	43.8	35.0	44.9	35.8	39.4	29.9
	BRCB	INT8	54.9	45.4	49.4	39.6	50.5	40.5	45.0	34.6
BEVFusion	*Baseline*	FP32	71.4	68.5	57.2	53.0	58.6	54.2	51.5	45.3
	MiniViT	FP32	70.0	67.0	56.0	51.8	57.5	53.0	50.4	44.3
	UVC	FP32	68.7	65.4	55.1	50.6	56.5	51.8	49.6	43.3
	FQ-ViT	INT8	69.2	65.8	55.4	50.9	56.8	52.1	49.9	43.5
	PTQ-ViT	INT8	69.5	66.1	55.6	51.1	57.0	52.3	50.1	43.7
	GPUSQ-ViT	INT8	70.6	67.4	56.6	52.1	58.0	53.3	50.9	44.6
	BRCB	INT8	70.8	67.6	63.8	59.0	65.2	60.3	58.1	51.5

Table 1: Compare the ***BRCB*** with state-of-the-art compression methods on autonomous driving (3D object detection) task.

model compression, **BRCB** can compress them into a more compact form. For the corruptions-sensitive elements, **BRCB** fully utilizes the given corrupted training samples during their compression process.

3.2 Push the Limitation of Robustness Boundary

In prior model compression techniques, the dense model's behavior is used as the golden reference, and the target compressed model is designed to mimic the dense model's inference results and feature maps. Suppose the dense model makes the wrong inference result due to the perturbation caused by changing from benign to adversarial samples. In that case, the compressed model will have a high possibility of making the same mistake. This principle design paradigm limits the compressed model, which ***cannot exceed the robustness boundary*** of dense model when facing corruption.

BRCB breaks the robustness boundary in the corruption-sensitive model compression path. Instead of mimicking the dense model as the golden reference, it tries to solve this problem from the principle view. The main reason that the corrupted samples bring misleading behavior to the embodied AI model is that the normal feature maps (i.e., the output activation tensors), when feeding the model with the benign inputs, have the out-of-range or out-of-distribution outliers introduced by the corruption. So, the **BRCB** algorithm attaches a ***BRCB gate layer*** before each compression target layer, like the query, key, value, linear projection, and feed-forward layers in the transformer structure. The primary purpose of the ***BRCB gate layer*** is to randomly mask out some elements in the feature maps before feeding them into the compression target layers during the corruption-sensitive model compression tuning process. With such regulation during training, the final obtained compressed model can properly adjust when meeting the outliers in inference with the corrupted samples.

We can find that the ***BRCB gate*** introduced in the corruption-sensitive model compression process has no reliance on the original dense model as a golden reference. In contrast, it enhances the robustness of the corruption for the compressed model from the principle. In the following experiments, we can see that the **BRCB**-compressed models can even exceed the ***robustness boundary*** of the dense model.

4 Experiments and Results

For the experiments in this section, we choose PyTorch with version 1.9.0 as the framework to train all baseline and efficient neural models. All of the training experimental results are obtained with A100 GPU clusters. All the accuracy results reported for our proposed method use INT8 as the default data type. All the reference algorithms use the default data type provided in public repositories.

4.1 Compression Efficacy for Autonomous Driving

To evaluate the compression efficacy of **BRCB** and make the comparison with prior arts on the autonomous driving, BEVFormer & BEVFormer v2 [Yang *et al.*, 2023a][1] and BEVFusion [Liu *et al.*, 2023b][2] are chosen as the experiment target models for 3D object detection task. BEVFusion is chosen as the experiment target model for bird's-eye view (BEV) map segmentation task. For the state-of-the-art transformer-based compression methods, we choose the MiniViT [Zhang *et al.*, 2022] and UVC [Yu *et al.*, 2022] as the reference sparse pruning techniques, we choose the FQ-ViT [Lin *et al.*, 2022] and PTQ-ViT [Liu *et al.*, 2021b] as the reference quantization techniques. We also choose the GPUSQ-ViT [Yu *et al.*,

[1]https://github.com/fundamentalvision/BEVFormer

[2]https://github.com/mit-han-lab/bevfusion

2023b] as the reference mixed-compression technique. The comparison results for **3D object detection** task and **BEV map segmentation** task are shown in Tables 1 and 2.

Model	Method	Format	Benign mIoU	Rainy mIoU	Snowy mIoU	Dark mIoU
BEVFusion	*Baseline*	FP32	63.0	50.4	51.7	45.4
	MiniViT	FP32 (Sparse)	61.1	48.9	50.1	44.0
	UVC	FP32 (Sparse)	60.0	48.1	49.3	43.3
	FQ-ViT	INT8	60.4	48.4	49.6	43.5
	PTQ-ViT	INT8	60.6	48.6	49.8	43.7
	GPUSQ-ViT	INT8 (Sparse)	61.9	49.6	50.9	44.7
	BRCB	INT8 (Sparse)	62.1	56.0	57.2	51.0

Table 2: Compare the ***BRCB*** with state-of-the-art compression methods on autonomous driving (BEV map segmentation) task.

The comparison results shown in Tables 1 and 2 show that **BRCB** can steadily provide a smaller accuracy drop for the compressed models on benign samples than both sparse pruning and quantization prior arts. And the accuracy drop compared with the original dense baseline model is negligible in most cases. There is an apparent accuracy drop when natural corruption like rain, snow, and darkness exists for the original dense models. Not to mention, the compressed models were obtained using the prior model compression techniques. However, the **BRCB** compressed counterparts have an apparent lower accuracy drop than other compressed models, even against the original dense model. This phenomenon proves the effectiveness of the against-corruption mechanism designed in the **BRCB** method. Moreover, it confirms that the compressed model can ***break the robustness boundary*** of the dense model.

Model	Method	Tasks / Attack Strength / Metrics	3D Object Detection 1% NDS	3D Object Detection 1% mAP	3D Object Detection 10% NDS	3D Object Detection 10% mAP	Segmentation 1% mIoU	Segmentation 10% mIoU
BEVFusion	*Baseline*	FP32	52.2	46.3	41.8	35.8	46.1	37.8
	MiniViT	FP32 (Sparse)	51.1	45.2	41.0	35.0	44.7	36.7
	UVC	FP32 (Sparse)	50.3	44.2	40.3	34.2	43.9	36.1
	FQ-ViT	INT8	50.6	44.5	40.5	34.4	44.2	36.3
	PTQ-ViT	INT8	50.8	44.6	40.7	34.5	44.4	36.4
	GPUSQ-ViT	INT8 (Sparse)	51.6	45.5	41.4	35.2	45.3	37.2
	BRCB	INT8 (Sparse)	58.9	52.4	47.2	40.5	51.7	42.4

Table 3: Compare the ***BRCB*** with state-of-the-art compression methods on autonomous driving task with adversarial corruption.

In Table 3, we also evaluate the effectiveness of the **BRCB** compression technique and compare it with the prior arts when facing adversarial corruption. Because adversarial corruption is manually intended to perturb the judgment of the embodied AI models, we can find the accuracy drops are more evident than facing natural corruption. The comparison results shown in Table 3 show that the **BRCB** compressed models can still steadily provide a smaller accuracy drop for the compressed models than prior arts, even against the original dense model. This experiment further proves the effectiveness of the against-corruption mechanism designed in the **BRCB** method.

4.2 Compression Efficacy for Robotics

To evaluate the compression efficacy of **BRCB** on the robotics, OpenVLA [Kim *et al.*, 2024][3] is chosen as the experiment target model for various tasks. For the state-of-the-art transformer-based compression method, we choose the GPUSQ-ViT [Yu *et al.*, 2023b]. The comparison results for **Robotics** task are shown in Tables 4.

We evaluate on a comprehensive set of evaluation tasks in each environment that covers various axes of generalization, such as visual (unseen backgrounds, distractor objects, colors/appearances of objects); motion (unseen object positions/orientations); physical (unseen object sizes/shapes); and semantic (unseen target objects, instructions, and concepts from the Internet) generalization. The comparison results shown in Table 4 show that the **BRCB** compressed models can still steadily provide better accuracy than the prior arts compressed models. And when facing the unseen situations, the generalization capabilities are even against the original dense model. This experiment further proves the effectiveness of the prposed **BRCB** method.

4.3 Visualization Experiments

For autonomous vehicle, real tasks are multidimensional and more complicated. We also verify **BRCB** workflow and optimizing the deployment efficiency of UniverseNet [Shinya, 2021] on object detection task, CondLaneNet [Liu *et al.*, 2021a] on lane detection task, Mask2Former [Cheng *et al.*, 2022] on segmentation task, and StereoDNN [Smolyanskiy *et al.*, 2018] on depth estimation task.

For the deployment efficiency, the **BRCB**-compressed UniverseNet model, the **BRCB**-compressed CondLaneNet model, the **BRCB**-compressed Mask2Former model and the **BRCB**-compressed StereoDNN model can achieve $2.31\times$, $2.55\times$, $1.97\times$ and $1.66\times$ acceleration against their corresponding dense counterparts on NVIDIA DRIVE AGX Orin platform. The corresponding visualization results are shown in Figure 5, Figure 6, and Figure 7. We can find all these autonomous driving models keep their effectiveness while the deployment efficiency is obviously improved.

5 Conclusions

In this paper, we have introduced a novel approach to enhance the efficiency and robustness of embodied AI models through the Boost Robust Compression Boundary (**BRCB**) algorithm. Our findings demonstrate that traditional model compression techniques often fall short when applied to the unique challenges faced by embodied AI systems, particularly in scenarios requiring rapid responses and high reliability.

The **BRCB** algorithm addresses these shortcomings by employing an innovative against-corruption mechanism that distinguishes between corruptions-stable and corruptions-sensitive model components. This fine-grained approach allows for more effective compression while maintaining the integrity of model performance across diverse conditions. Our experimental results indicate that models compressed using **BRCB** not only achieve comparable accuracy on benign inputs but also exhibit significantly improved resilience against various natural and adversarial corruptions.

Moreover, the **BRCB** method has shown a remarkable ability to push beyond the robustness boundaries established by dense models, enabling compressed models to perform better

[3]https://github.com/openvla/openvla

Model	Method	Format	Visual Generalization Success Rate(%)	Motion Generalization Success Rate(%)	Physical Generalization Success Rate(%)	Sementatic Generalization Success Rate(%)
OpenVLA	*Baseline*	FP32	87.0	60.0	76.7	36.3
	GPUSQ-ViT	INT8	79.5	53.6	67.7	29.8
	BRCB	INT8	89.1	61.0	77.5	36.7

Table 4: Compare the ***BRCB*** with state-of-the-art compression method on robotics task.

Figure 5: **(Left)** Visualization results of compressed UniverseNet model in object detection and tracking tasks. (The six images are captured with the surrounded cameras on the autonomous car.) **(Right)** Visualization results of compressed CondLaneNet model in lane detection task.

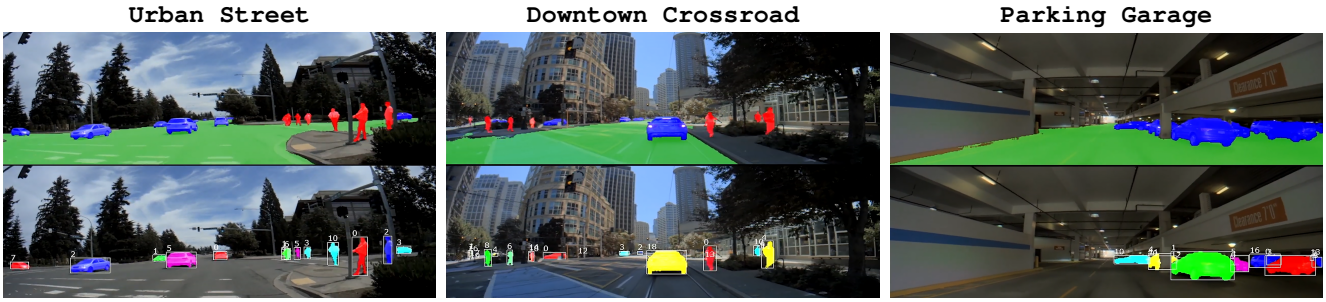

Figure 6: Visualization for semantic object segmentation and instance segmentation results from the compressed Mask2Former model. (The upper row shows the semantic object segmentation results. The lower row shows the instance segmentation results.)

Figure 7: Depth estimation results by visual odometer model (the compressed StereoDNN model) deployed on NVIDIA DRIVE AGX Orin.

than their uncompressed counterparts in challenging environments. This advancement is particularly relevant for applications in autonomous driving and robotics, where safety and reliability are paramount. We also evaluate the efficacy of our method on BEV, object detection, lane detection, segmentation, and depth estimation tasks with the real deployment efficiency on the DRIVE AGX Orin autonomous vehicle platform.

Ethical Statement

This research does not involve human or animal subjects, and therefore does not present any ethical issues related to participant consent or welfare. The methodologies employed in this study focus solely on computational models and algorithms, ensuring compliance with ethical standards in artificial intelligence research.

Acknowledgements

This work is supported by Shanghai Natural Science Foundation (No. 23ZR1402900), National Key Research and Development Program of China (No. 2022ZD0160101), Shanghai Science and Technology Commission Explorer Program Project (24TS1401300), Shanghai Municipal Science and Technology Major Project (No.2021SHZDZX0103).

The computations in this research were performed using the CFFF platform of Fudan University.

References

[Abbasi *et al.*, 2022] Rashid Abbasi, Ali Kashif Bashir, Hasan J Alyamani, Farhan Amin, Jaehyeok Doh, and Jianwen Chen. Lidar point cloud compression, processing and learning for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 24(1):962–979, 2022.

[Brohan *et al.*, 2022] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[Chen *et al.*, 2018] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018.

[Cheng *et al.*, 2022] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022.

[Chi *et al.*, 2023] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023.

[Diffenderfer *et al.*, 2021] James Diffenderfer, Brian Bartoldson, Shreya Chaganti, Jize Zhang, and Bhavya Kailkhura. A winning hand: Compressing deep networks can improve out-of-distribution robustness. *Advances in neural information processing systems*, 34:664–676, 2021.

[Gui *et al.*, 2019] Shupeng Gui, Haotao Wang, Haichuan Yang, Chen Yu, Zhangyang Wang, and Ji Liu. Model compression with adversarial robustness: A unified optimization framework. *Advances in Neural Information Processing Systems*, 32, 2019.

[Hnewa and Radha, 2020] Mazin Hnewa and Hayder Radha. Object detection under rainy conditions for autonomous vehicles: A review of state-of-the-art and emerging techniques. *IEEE Signal Processing Magazine*, 38(1):53–67, 2020.

[Jiang *et al.*, 2023] Yanqin Jiang, Li Zhang, Zhenwei Miao, Xiatian Zhu, Jin Gao, Weiming Hu, and Yu-Gang Jiang. Polarformer: Multi-camera 3d object detection with polar transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1042–1050, 2023.

[Khazatsky *et al.*, 2024] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

[Kim *et al.*, 2024] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[Laurençon *et al.*, 2022] Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. The bigscience roots corpus: A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems*, 35:31809–31826, 2022.

[Li *et al.*, 2022] Jiale Li, Hang Dai, and Yong Ding. Self-distillation for robust lidar semantic segmentation in autonomous driving. In *European Conference on Computer Vision*, pages 659–676. Springer, 2022.

[Li *et al.*, 2023] Jianing Li, Ming Lu, Jiaming Liu, Yandong Guo, Yuan Du, Li Du, and Shanghang Zhang. Bev-lgkd: A unified lidar-guided knowledge distillation framework for multi-view bev 3d object detection. *IEEE Transactions on Intelligent Vehicles*, 2023.

[Lin *et al.*, 2022] Yang Lin, Tianyu Zhang, Peiqin Sun, Zheng Li, and Shuchang Zhou. Fq-vit: Post-training quantization for fully quantized vision transformer. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 1173–1179, 2022.

[Liu *et al.*, 2021a] Lizhe Liu, Xiaohao Chen, Siyu Zhu, and Ping Tan. Condlanenet: a top-to-down lane detection framework based on conditional convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3773–3782, 2021.

[Liu *et al.*, 2021b] Zhenhua Liu, Yunhe Wang, Kai Han, Wei Zhang, Siwei Ma, and Wen Gao. Post-training quantization for vision transformer. *Advances in Neural Information Processing Systems*, 34:28092–28103, 2021.

[Liu *et al.*, 2023a] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Aqi Gao, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petrv2: A unified framework for 3d perception from multi-camera images. In *Proceedings of the IEEE International Conference on Computer Vision*, 2023.

[Liu *et al.*, 2023b] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2774–2781. IEEE, 2023.

[Luo *et al.*, 2022] Rui Luo, Chunpeng Wang, Eric Schwarm, Colin Keil, Evelyn Mendoza, Pushyami Kaveti, Stephen Alt, Hanumant Singh, TaŞkin Padir, and John Peter Whitney. Towards robot avatars: Systems and methods for teleinteraction at avatar xprize semi-finals. In *2022 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 7726–7733. IEEE, 2022.

[Mees *et al.*, 2024] Oier Mees, Dibya Ghosh, Karl Pertsch, Kevin Black, Homer Rich Walke, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, et al. Octo: An open-source generalist robot policy. In *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024.

[Shen *et al.*, 2021] Yu Shen, Laura Zheng, Manli Shu, Weizi Li, Tom Goldstein, and Ming Lin. Gradient-free adversarial training against image corruption for learning-based steering. *Advances in Neural Information Processing Systems*, 34:26250–26263, 2021.

[Shinya, 2021] Yosuke Shinya. Usb: Universal-scale object detection benchmark. *arXiv preprint arXiv:2103.14027*, 2021.

[Smolyanskiy *et al.*, 2018] Nikolai Smolyanskiy, Alexey Kamenev, and Stan Birchfield. On the importance of stereo for accurate depth estimation: An efficient semi-supervised deep neural network approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1007–1015, 2018.

[Tinchev *et al.*, 2019] Georgi Tinchev, Adrian Penate-Sanchez, and Maurice Fallon. Learning to see the wood for the trees: Deep laser localization in urban and natural environments on a cpu. *IEEE Robotics and Automation Letters*, 4(2):1327–1334, 2019.

[Touvron *et al.*, 2023] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[Walke *et al.*, 2023] Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pages 1723–1736. PMLR, 2023.

[Wan *et al.*, 2021] Zishen Wan, Bo Yu, Thomas Yuang Li, Jie Tang, Yuhao Zhu, Yu Wang, Arijit Raychowdhury, and Shaoshan Liu. A survey of fpga-based robotic computing. *IEEE Circuits and Systems Magazine*, 21(2):48–74, 2021.

[Wang *et al.*, 2019] Zhiyong Wang, Jinxiang Chai, and Shihong Xia. Combining recurrent neural networks and adversarial training for human motion synthesis and control. *IEEE transactions on visualization and computer graphics*, 27(1):14–28, 2019.

[Wang *et al.*, 2023] Shunxin Wang, Raymond Veldhuis, Christoph Brune, and Nicola Strisciuglio. A survey on the robustness of computer vision models against common corruptions. *arXiv preprint arXiv:2305.06024*, 2023.

[Xie *et al.*, 2024] Annie Xie, Lisa Lee, Ted Xiao, and Chelsea Finn. Decomposing the generalization gap in imitation learning for visual robotic manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3153–3160. IEEE, 2024.

[Yan *et al.*, 2018] Ziang Yan, Yiwen Guo, and Changshui Zhang. Deep defense: Training dnns with improved adversarial robustness. *Advances in Neural Information Processing Systems*, 31, 2018.

[Yang *et al.*, 2023a] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, et al. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17830–17839, 2023.

[Yang *et al.*, 2023b] Jing Yang, Nade Liang, Brandon J Pitts, Kwaku O Prakah-Asante, Reates Curry, Mike Blommer, Radhakrishnan Swaminathan, and Denny Yu. Multimodal sensing and computational intelligence for situation awareness classification in autonomous driving. *IEEE Transactions on Human-Machine Systems*, 53(2):270–281, 2023.

[Yu *et al.*, 2022] Shixing Yu, Tianlong Chen, Jiayi Shen, Huan Yuan, Jianchao Tan, Sen Yang, Ji Liu, and Zhangyang Wang. Unified visual transformer compression. In *International Conference on Learning Representations*, 2022.

[Yu *et al.*, 2023a] Chong Yu, Tao Chen, and Zhongxue Gan. Adversarial amendment is the only force capable of transforming an enemy into a friend. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 4522–4530, 2023.

[Yu *et al.*, 2023b] Chong Yu, Tao Chen, Zhongxue Gan, and Jiayuan Fan. Boost vision transformer with gpu-friendly sparsity and quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22658–22668, 2023.

[Yu, 2021] Chong Yu. Minimally invasive surgery for sparse neural networks in contrastive manner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3589–3598, 2021.

[Zhang *et al.*, 2022] Jinnian Zhang, Houwen Peng, Kan Wu, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. Minivit: Compressing vision transformers with weight multiplexing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12145–12154, 2022.