CrossBack: Selective Knowledge Transfer for Cross-modal devices in Federated Continual Learning

Minju Cho[†1], Jisoo Jeong[†1] and YouLee Han[†1]

[1] Ewha Womans University/Computer Science & Engineering, Seoul, Republic of Korea, {mummyee, cjslove0530, oneglass}@ewhain.net

Abstract

Multimodal artificial intelligence (AI) has gained prominence for its ability to deliver inference accuracy and robustness unattainable by unimodal approaches. However, deploying such models on low-end edge devices is hampered by stringent constraints on computation, memory, security, and power, as well as by the pronounced heterogeneity of input resolutions, sequence lengths, and sensor densities across devices. These factors render uniform model distribution and global update strategies ineffective.

We introduce selective knowledge transfer, whereby each device solely assimilates external heterogeneous knowledge most beneficial to its local context. *CrossBack* extends the original Pick-a-back paradigm beyond its single-modality, fixed-input assumptions: it quantifies decision-pattern similarity across differing modalities and input scales, enabling each model to reuse the backbone of its most compatible peer without reliance on a central server.

Experiments on CUB-200-2011, and Oxford-102 Flowers demonstrate that models trained on orthogonal modalities-such as images and text-can engage in effective selective knowledge transfer via *CrossBack*. This process simultaneously increases accuracy and reduces computational overhead, validating the suitability of the framework for multimodal real-time collaboration on resource-constrained edge devices. *CrossBack* thus advances edge AI beyond single-modality assumptions, establishing a foundation for continual, autonomous learning in decentralized multimodal systems.

Keywords— *Continual Learning, Knowledge Transfer, Decentralized Learning, Multimodal AI, Cross-Modality, Deep Learning*

[1]Source code available at `https://github.com/EWHA-Tespa/CrossBack`

[†]These authors contributed equally to this work.

I. INTRODUCTION

With the rapid advancement of artificial intelligence (AI), multimodal AI—which integrates data generated from diverse modalities, such as visual, auditory, textual, and sensor streams—has emerged as a key focus of research and applications. Synergistic interactions among modalities can deliver inference accuracy and robustness unattainable by unimodal approaches, making the efficient exploitation of multimodal data a central challenge in the current AI era.

Concurrently, demand is growing for deploying multimodal processing on edge devices that compute locally at the data source, rather than relying on centralized cloud infrastructure. By processing data where it is generated, edge AI reduces latency, enhances privacy, and lowers communication costs. Therefore, combining multimodal processing with edge deployment is pivotal for real-world scenarios ranging from autonomous vehicles and smart manufacturing to wearable technologies.

In practice, however, most deployed systems rely on low-end edge devices with stringent limits on computation, memory, and power budgets. These resource constraints make the on-device deployment of large-scale multimodal models challenging. In addition, the data collected on each device may differ markedly in resolution, sequence length, and sensor density, leading to pronounced input-size mismatches. This cross-device heterogeneity remains an unresolved challenge, particularly in federated learning settings with heterogeneous, decentralized multimodal data sources.

Human intelligence seamlessly integrates information from multiple senses and only shares the knowledge relevant to the current context. Likewise, edge AI systems must flexibly share information despite limited resources. In this work, we define selective knowledge transfer as a mechanism that allows multimodal models—with various modalities and model capacities—to exchange mutually beneficial knowledge. In resource-constrained edge settings, data on low-end devices often varies widely in such heterogeneity. Consequently, deploying a uniform model or pushing identical updates to every node is frequently inefficient—if not infeasible. Each local model should there-

fore selectively absorb external knowledge most relevant to its context, making selective knowledge transfer indispensable.

A notable step toward decentralized continual learning is Pick-a-back [17], which enables device-to-device knowledge reuse by allowing edge models to selectively incorporate the backbone of a peer with similar decision patterns. This approach removes the need for a centralized server and supports continual adaptation through local knowledge assimilation. However, Pick-a-back assumes homogeneous input modalities and uniform input sizes across devices, limiting its scalability in real-world edge environments characterized by multimodal and heterogeneous data. Our work builds upon this framework by extending it to handle input- and modality-level diversity, enabling cross-modal, input-scale-aware knowledge transfer without the need for centralized coordination.

Motivated by these limitations, we pose the following overarching question: How can models with heterogeneous input sizes and modalities efficiently leverage external knowledge? Answering it is critical for moving beyond single modality assumptions toward knowledge transfer strategies that operate under realistic edge constraints. We therefore decompose the problem into two research thrusts.

- Cross-size knowledge transfer: How can models trained on inputs that differ in spatial or temporal resolution share knowledge without sacrificing critical detail? Naïve resizing often erodes fine-grained cues, and uniform normalization is infeasible when each edge device operates under distinct sensing conditions. We thus seek mechanisms that allow models to preserve their own inductive biases while still absorbing practical knowledge across heterogeneous input sizes.

- Cross-modal utility estimation: When representation spaces are heterogeneous—for example, vision versus language—similarity-based selection becomes unreliable. How can we quantify compatibility across modalities and use that metric to decide which external representations are worth transferring? Establishing principled measures of cross-modal alignment is essential for effective selective knowledge transfer in multimodal edge deployments.

To address such challenges, we propose *CrossBack*, a unified framework for selective multimodal knowledge transfer on edge devices. Our approach extends the Pick-a-back paradigm to handle input-heterogeneous and modality-diverse scenarios, without relying on centralized servers or modality-specific supervision. Our main contributions are summarized as follows:

- Unified multimodal edge knowledge transfer: We formulate the challenge of device-to-device knowledge transfer under modality and input-scale heterogeneity, combining disparate input formats and decentralized learning into a single coherent problem.

- *CrossBack* framework for cross-modal continual knowledge transfer: We extend the Pick-a-back [17] paradigm to multimodal settings by allowing models to evaluate decision pattern similarity even across modalities. Using *CrossBack*, each model identifies and reuses the backbone of the most compatible peer, even when trained on a different modality. This selective reuse not only supports modality-agnostic continual federated learning but also improves performance through cross-modal knowledge complementation.

- Empirical validation: Experiments on two benchmarks: CUB-200-2011 [16], and Oxford-102 Flowers [12] demonstrate that models trained independently on images and text can engage in selective knowledge transfer, improving accuracy and reducing computational cost even under severe size- and modality-mismatch conditions.

II. Related Work

Multimodal Learning. Multimodal learning seeks to integrate heterogeneous sources such as images, text, and audio to build models with richer contextual understanding. Early approaches leveraged indirect semantic supervision, where image models were enhanced through co-occurring textual data. For example, Mori et al [11]. and Quattoni et al [13]. demonstrated that predicting descriptive words from associated documents or captions could improve visual representations. However, these methods relied on paired training data and assumed aligned modality availability, making them unsuitable for edge scenarios where modalities may not co-occur.

To mitigate this dependency, later works shifted toward joint representation learning using deep generative models. Srivastava and Salakhutdinov's multimodal Deep Boltzmann Machines [15] and Frome et al.'s DeViSE projected features into a shared latent space for cross-modal alignment [1]. These models assumed access to modality-aligned training distributions and full data centralization—assumptions incompatible with heterogeneous, decentralized edge deployments where data resides on-device and input formats vary.

Transformer-based architecture like CLIP[14] has further advanced contrastive multimodal pretraining by learning alignment across large-scale image-text pairs. Recently, C-CLIP[7], a benchmark model for multimodal continual learning, has been proposed to extend the CLIP framework. C-CLIP demonstrated the ability to retain prior knowledge while maintaining initial performance over time. Trained on diverse domain-specific image-text datasets, it achieved state-of-the-art performance, outperforming existing vision-language models and offer-

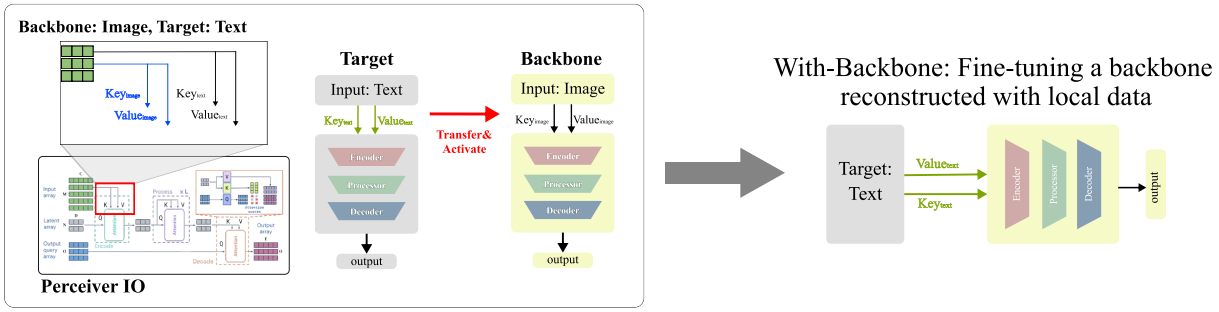

Fig. 1. Overview of *CrossBack*'s cross-modal selective knowledge transfer process. The left side illustrates the backbone model trained on image modality, while the right shows the target model using text input. Key/value projections from the image-based Perceiver IO backbone are transferred and activated in the text-based model (Transfer & Activate). A new PiggyMask is applied to the transferred layers to enable selective adaptation. The reconstructed backbone is then fine-tuned using local text data, allowing continual learning without relying on a central server.

ing significant advances in continual multimodal learning. While effective, their methods require massive centralized datasets with uniform input preprocessing pipelines, and cannot flexibly adapt to devices that differ in resolution, sequence length, or input modality. Moreover, they lack mechanisms for peer-to-peer knowledge sharing, which are critical in resource-constrained edge intelligence scenarios.

In contrast, our method, *CrossBack*, explicitly targets decentralized and heterogeneous settings. Rather than assuming globally shared modalities or centralized training, we enable modality-agnostic and scale-agnostic knowledge transfer between devices via internal representation alignment, filling a fundamental gap left by prior multimodal methods.

Federated and Continual Learning. Federated learning (FL) enables model training across distributed clients without data sharing. Foundational methods like FedAvg [10] and FedProx [5] rely on server-based aggregation and assume task and model homogeneity, which limits their applicability in real-world edge scenarios where both input modalities and computational capabilities are diverse. These approaches also fail to address continual adaptation, leading to performance degradation in non-stationary environments.

To solve such shortcomings, federated continual learning (FCL) has emerged, with many works addressing class-incremental learning in server-coordinated environments. However, these approaches—while effective in controlled settings—overlook modality heterogeneity and typically fail to accommodate devices with fundamentally different data types and model constraints.

Parameter-efficient methods such as Piggyback [8] and CPG [2] introduced task-specific masking strategies to avoid catastrophic forgetting, enabling continual learning with minimal parameter overhead. Yet, they presuppose that a shared backbone exists and is applicable across all tasks, which breaks down in cross-modal contexts where input structures and latent representations diverge.

Building upon these masking-based strategies, Pick-a-back introduced a decentralized framework for knowledge reuse, in which each device selectively adopts a neighbor's backbone based on decision-pattern similarity. However, the original method assumes that all participating devices operate under identical input modalities and structural configurations, which restricts its applicability to unimodal and resolution-consistent scenarios. As a result, it fails to support knowledge transfer across structurally heterogeneous devices commonly found in real-world edge environments.

To overcome these limitations, we propose *CrossBack*, an extension of the Pick-a-back framework designed to accommodate cross-modal and cross-scale heterogeneity. This design enables a novel capability: modality-agnostic backbone reuse, which has not been demonstrated in any prior decentralized continual learning approach. In our experiments, we include the original Pick-a-back as a baseline to quantitatively evaluate how mismatches in modality and input scale affect transferability. The results demonstrate that *CrossBack* significantly improves knowledge sharing performance in heterogeneous environments.

III. Approach

In real-world edge AI environments, deployed systems often encounter diverse data distributions in both modality (e.g., image vs. text) and shape (e.g., varying sequence lengths, resolutions, or feature dimensions). Traditional continual learning frameworks assume homogeneous inputs, which limits their applicability in unimodal environments. To address this, we propose a modality-agnostic continual federated learning framework that supports selective knowledge transfer across heterogeneous edge devices. This framework, referred to as *CrossBack*, extends the Pick-a-back architecture by enabling cross-modal, shape-aware knowledge transfer between edge models without requiring centralized orchestration.

Algorithm 1. **CrossBack**

Require: $L, \{M_0^{(l)}, T^{(l)}, \theta^{(l)}, CL, \text{piggyMask}\}_{l=1}^{L}$
for $l = 1, \ldots, L$ **do**
  for $i = 1, \ldots, |T^{(l)}|$ **do**
    $M_{\text{loc}} \leftarrow CL.\text{train}(M_{i-1}^{(l)}, T_i^{(l)}, \theta_i^{(l)})$ ▷ local learning
    for $k = 1, \ldots, L$ **do**
      $\text{sim}[k] \leftarrow$
    $\text{combine}\big(cos\_ddv\big(\ DDV(\text{attnRep}(M_{\text{loc}}, \text{extractKV}(M_{i-1}^{(k)}))),$
    $DDV(\text{attnRep}(M_{i-1}^{(k)}, \text{extractKV}(M_{i-1}^{(k)})))\big),\ euc\_ddv(\ldots)\big)$ ▷ attention
    end for
    $k^* \leftarrow \arg\max_k \text{sim}[k]$
    $\hat{M} \leftarrow \text{init\_piggyMask}(M_{\text{loc}}, \text{extractKV}(M_{i-1}^{(k^*)}))$ ▷ fuse KV
    $M_i^{(l)} \leftarrow CL.\text{train\_with\_mask}(\hat{M}, T_i^{(l)}, \theta_i^{(l)})$ ▷ masked adaptation
  end for
end for
return $\{M_{|T^{(l)}|}^{(l)}\}_{l=1}^{L}$

Notation. L is the number of devices; $M_0^{(l)}$ initial Perceiver IO on device l; $T^{(l)}$ its task sequence; $\theta^{(l)}$ task hyperparameters; CL the continual-learning strategy; init_piggyMask injects a piggyMask over extracted KV layers; extractKV(M) extracts the key–value pair from M; attnRep(M, kv) returns the cross-attention output representation of model M given key–value input kv; cos_ddv, euc_ddv compute cosine- and Euclidean-based distributional distances of DDVs.0

A. Cross-Modal Selective Knowledge

The proposed framework operates in a decentralized federated learning environment, where each decentralized edge device possesses either image or text modal local data. Each device learns in a continual, task-specific manner, encountering a sequence of local tasks while maintaining compatibility with potential knowledge sources from other devices. The principal elements of the framework are as follows:

- **Backbone Selection**: For each local task, the framework identifies the most compatible peer model as a backbone. This selection is performed by comparing model responses in a latent representation space *induced by cross-attention interactions*, using shared and perturbed key/value projections extracted from each model.

- **Cross-modal and cross-shape compatibility**: The framework Supports knowledge transfer across both modalities (e.g., image vs. text) and input dimensionality gaps. Instead of aligning raw inputs or architectural structures, it operates on internal attention-based representations that are normalized, mutated, and compared using distance-based metrics.

- **Continual Adaptation**: Once the backbone model is selected, its internal key/value projection layers are transferred to the local model. The local model then continues learning using the CPG [2] paradigm, which mitigates catastrophic forgetting by dynamically reallocating parameter capacity.

IV. Architectures

To evaluate the effectiveness of cross-modal knowledge transfer in continual federated learning, we implemented our architecture using lightweight and modality-flexible architectures. Our implementation is composed of two core modules: the continual learner and the federated knowledge transfer agent.

A. Continual Learner

Each local device is equipped with a Perceiver IO [3] model as its backbone. The model is chosen for its architectural flexibility to handle diverse input modalities (e.g., text sequences and image patches) with varying shapes. Following the original Pick-a-back, we employ two representative continual learning strategies to mitigate catastrophic forgetting during sequential task training: PackNet [9] and CPG [2]. It manages model capacity during sequential task learning to allow each model to adapt efficiently to new tasks while avoiding catastrophic forgetting at the same time.

B. Multimodal Federated Learner

The federated component of our system operates without a central server. When a new task arrives at a local device, it computes task similarity to other devices' models using a variation of ModelDiff [6]. We first extracted key and value projection layers from the cross-attention to compare models with different input formats. Then, we compute the decision distance vector (DDV) for each model, which is a metric that measures decision pattern similarity for the model output pair. We compute the DDV-cosine and DDV-euclidean distances between devices, thereby choosing the most compatible model as the backbone for knowledge transfer. The selected backbone model may come from either the same or a different modality. Once the device chooses its backbone provider, its key/value layers are transferred to the local model and integrated with a newly initialized piggymask, distinct from previously learned masks, allowing robust adaptation of injected layers in the fine-tuning step.

V. Experiments

A. Experimental Design

To evaluate the effectiveness of *CrossBack* under realistic, continual federated learning scenarios, we designed a set of experiments simulating decentralized edge environments with heterogeneous devices. Each device is assigned a sequence of tasks, either in image or text modality, and engages in continual learning over time without access to a centralized server or shared raw data. To evaluate cross-modal compatibility, we designed experiments using multimodal datasets with hierarchical class structures.

Table 1. Compatibility Evaluation Results: Comparison of accuracy between `wo_backbone` and `w_backbone` across tasks on two datasets.

(a) CIFAR-100 (Image modality with Perceiver IO)

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Selected Backbone ID	5	3	12	5	12	3	5	3	3	3	5	3	3	5	3	3	3	5	18	18
Wo_Backbone	0.516	0.658	0.568	0.444	0.672	0.446	0.530	0.538	0.514	0.546	0.740	0.464	0.514	0.480	0.318	0.438	0.454	0.492	0.442	0.714
W_Backbone	**0.546**	0.636	**0.588**	**0.514**	**0.734**	**0.524**	**0.548**	**0.650**	**0.694**	**0.634**	0.692	**0.556**	**0.646**	**0.598**	**0.396**	**0.474**	**0.538**	**0.536**	**0.584**	**0.720**

(b) Oxford 102 Flowers (Tasks 1–6: Image, Tasks 7–12: Text)

ID	1	2	3	4	5	6	7	8	9	10	11	12
Selected Backbone ID	5	5	5	5	5	5	11	7	11	8	8	7
Wo_Backbone	0.8188	0.9661	0.8389	0.5932	0.8974	0.6695	0.6411	0.9059	0.7966	0.5085	0.8633	0.6271
W_Backbone	0.7778	0.9573	**0.8898**	**0.7966**	**0.9316**	**0.7542**	**0.7009**	**0.9145**	0.7373	**0.5169**	**0.8803**	0.5763

We used four different groups of datasets: CIFAR-100 [4], CUB-200-2011 [16], and Oxford 102 Flowers [12]. All but CIFAR-100 were preprocessed as follows: image and text data were separated, and trained to modality-specific Perceiver IO models separately. Half of the tasks were assigned to text-based models and the other half to image-based models, thus simulating modal heterogeneity.

Architecture. Our full experimental pipeline consists of five stages, each corresponding to a critical phase in continual learning and selective knowledge transfer. We used Perceiver IO implemented with PyTorch for our main evaluation.

Evaluation Metrics. Final task accuracy was used as the primary evaluation metric to assess model performance and forgetting. To validate the backbone selection mechanism and quantify transfer effectiveness from cross-modal backbones versus unimodal peers, we used the similarity distribution of DDV-cosine and DDV-euclidean.

Baselines. A modality-specific baseline is first established for each edge device, with half of the tasks trained on text-based Perceiver IO, while the other half on image-based models. Each device sees tasks sequentially and independently, and all models are initialized without prior knowledge of each other's internal states. Such a setup isolates the learner without any external knowledge transfer, serving as the lower bound for performance comparison.

To validate the compatibility of our framework with prior work and ensure consistent experimental grounding, we first tested our approach with the CIFAR-100 dataset. As a unimodal image classification benchmark previously used in the original Pick-a-back, CIFAR-100 allows us to verify that our Perceiver IO-based implementation performs comparably in single-modality continual learning settings.

We then extended our evaluation to the Oxford 102 Flowers dataset to validate whether the original Pick-a-back—designed for unimodal image classification—could be effectively applied to the text modality and intra-modal knowledge transfer within *CrossBack*.

To evaluate the full cross-modal transfer capability of *CrossBack*, we used a richly structured multimodal dataset: CUB-200-2011. It is reorganized into superclass–subclass hierarchies to simulate realistic continual learning tasks with diverse semantics across modalities.

In this phase, we designed two experimental conditions:

- **Intra-modal only:** Backbone candidates are selected exclusively from models trained in the same modality (image-to-image or text-to-text).
- **Cross-modal allowed:** Backbone candidates can be drawn from the full pool of models, regardless of modality, enabling backbone search across diverse modalities.

By comparing performance across these two conditions, we assess the effectiveness of *CrossBack* in selecting appropriate backbone models in both modality-aligned and modality-heterogeneous settings. This comparison reveals the framework's ability to generalize beyond modality boundaries in federated continual learning.

VI. Evaluation

A. Key result Overview

Through various experiments, we have demonstrated the effectiveness of the proposed selective backbone transfer strategy in both intra-modal and cross-modal environments. Specifically, the performance was evaluated from the following three perspectives:

1) Compatibility. To verify the compatibility of our approach with the original Pick-a-back method, we conducted continual learning experiments in a single modality: image-based tasks on the CIFAR-100 dataset and text-based tasks on the Oxford 102 Flowers dataset. In both experiments, the proposed *CrossBack* strategy either matched or outperformed the original approach, confirming that the Perceiver IO-based architecture is well-aligned with existing mechanisms.

Table 2. Intra-modal Transfer Results: Comparison of accuracy between `wo_backbone`, `random` and `w_backbone` for same-modality transfer settings using *CrossBack*.

(a) CUB-200-2011: Task IDs 1–28 (image), 29–56 (text)

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Selected Backbone ID	5	3	2	5	12	7	5	3	3	3	10	5	2	20
Wo_Backbone	0.4500	0.4667	0.4167	0.6333	0.6667	0.2667	0.5167	0.2500	0.4167	0.5167	0.1833	0.5833	0.3833	0.6333
Random	0.3500	0.4500	0.3500	0.6333	0.5333	0.3333	0.5333	0.2667	0.5000	0.6500	0.4500	0.6000	0.2833	0.5167
W_Backbone	**0.5833**	0.4000	**0.5333**	**0.8167**	**0.7667**	**0.4500**	**0.7333**	**0.3000**	0.3667	**0.6000**	**0.4500**	**0.7833**	0.3333	**0.6500**
ID	15	16	17	18	19	20	21	22	23	24	25	26	27	28
Selected Backbone ID	3	3	8	8	3	5	5	3	8	3	3	3	3	12
Wo_Backbone	0.4500	0.3167	0.2500	0.3167	0.2500	0.7000	0.2833	0.3667	0.2667	0.4667	0.3167	0.3000	0.4500	0.3000
Random	0.4333	0.3500	0.2833	0.2333	0.3167	0.6500	0.3333	0.3167	0.2000	0.3833	0.3000	0.2833	0.4000	0.3333
W_Backbone	**0.5333**	0.2500	0.2167	0.2667	**0.3833**	0.6833	**0.4667**	**0.3833**	0.2333	**0.5167**	**0.3333**	0.3000	0.2500	**0.5500**
ID	29	30	31	32	33	34	35	36	37	38	39	40	41	42
Selected Backbone ID	52	52	37	52	52	54	52	52	52	52	40	37	52	35
Wo_Backbone	0.5167	0.6500	0.5500	0.5333	0.5333	0.4833	0.7167	0.4000	0.4167	0.6500	0.4000	0.5500	0.4667	0.3667
Random	0.2167	0.6667	0.4167	0.5000	0.5500	0.3833	0.6667	0.2333	0.4333	0.2167	0.3667	0.5667	0.4833	0.3833
W_Backbone	**0.6333**	0.6000	0.4333	0.5167	**0.6167**	0.4500	0.6333	0.3333	**0.4833**	0.5667	0.3833	0.3500	0.4667	**0.6000**
ID	43	44	45	46	47	48	49	50	51	52	53	54	55	56
Selected Backbone ID	52	37	37	43	52	40	40	37	42	47	45	48	52	52
Wo_Backbone	0.3833	0.3333	0.2833	0.2667	0.5333	0.4833	0.1667	0.4333	0.5333	0.5000	0.3133	0.3167	0.3833	0.5500
Random	0.5000	0.2833	0.3000	0.2667	0.2000	0.6000	0.3333	0.3000	0.4833	0.3000	0.4000	0.2000	0.3500	0.4833
W_Backbone	**0.4833**	**0.4333**	**0.3000**	0.2667	0.4333	**0.5667**	**0.3000**	0.3833	0.4833	**0.5833**	0.2667	**0.4167**	**0.4500**	0.4167

2) Intra-modal Transfer. Experiments on backbone transfer within the same modality, using the CUB-200-2011 datasets, showed that the selected backbones maintain high accuracy within the same modality and transfer effectively. Notably, higher similarity based on DDV led to improved accuracy, further validating the effectiveness of selective transfer even in intra-modal settings.

3) Cross-modal Transfer. In experiments allowing cross-modal backbone selection on the CUB-200-2011 dataset, a significant number of tasks benefited from backbones selected across modalities. This demonstrated that even with modality mismatches, transfer across modalities based on distributional similarities is achievable. This suggests that *CrossBack* performs a harmonious transfer strategy that is both modality-aware and modality-agnostic.

4) Comparative Analysis of Backbone Usage: without backbone, random transfer, and with backbone. To further evaluate the effectiveness of the backbone selection strategy, we compared the performance of models under three different scenarios:

- **without backbone:** This scenario represents intact continual learning without using any selective knowledge transfer, essentially treating the model as a baseline.
- **random backbone:** This control setting introduces a randomly selected knowledge as a backbone for knowledge transfer.
- **with backbone:** This scenario includes the use of our strategy for backbone selection.

We compared the accuracy across each task ID to assess whether the backbone selection strategy contributes to performance improvement. The results confirmed that the use of *CrossBack* (`w_backbone`) provided significant performance gains compared to the baseline (`wo_backbone`) and the random backbone selection scenario (`random`). This highlights that performance improvement stems not merely from transferring information, but from selecting a semantically compatible backbone.

B. Experimental Results

As illustrated in Table 1, the *CrossBack* framework demonstrated performance either in alignment with or superior to that of the original approach, confirming that the Perceiver IO-based architecture and *CrossBack* framework are well-aligned with existing mechanisms. Performance was evaluated by comparing `wo_backbone` and `w_backbone`, using accuracy as the metric. The effectiveness of the backbone selection strategy was assessed through performance comparisons for each task ID.

CIFAR-100 Results. For most tasks, the model with the selected backbone (`w_backbone`) outperformed the model without the backbone (`wo_backbone`). Notable improvements were observed in tasks ID 5, 9, 12, 14, and 20, where clear performance gains were seen. These results demonstrate that the proposed method works effectively in an image-based single modality environment. The experiments, conducted using the Perceiver IO architecture, confirm the validity of the backbone selection strategy across various tasks in the image modality.

Table 3. Cross-modal Transfer Results on the CUB-200-2011 Dataset. Each task selects a backbone across modalities; comparison of `wo_backbone` vs. `w_backbone`, with improvements highlighted.

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Selected Backbone ID	35	31	49	38	55	30	40	53	46	45	31	53	35	35
Wo_Backbone	0.4500	0.4667	0.4167	0.6333	0.6667	0.2667	0.5167	0.2500	0.4167	0.5167	0.1833	0.5833	0.3833	0.6333
W_Backbone	**0.4833**	**0.5500**	**0.4667**	**0.7500**	**0.7167**	**0.4833**	0.4833	**0.3333**	0.3833	**0.5667**	**0.4167**	0.5167	**0.4167**	0.5333
ID	15	16	17	18	19	20	21	22	23	24	25	26	27	28
Selected Backbone ID	38	42	32	53	38	39	47	38	37	32	50	32	46	44
Wo_Backbone	0.4500	0.3167	0.2500	0.3167	0.2500	0.7000	0.2833	0.3667	0.2667	0.4667	0.3167	0.3000	0.4500	0.3000
W_Backbone	**0.4833**	**0.4333**	**0.2833**	0.3000	**0.5667**	0.6000	**0.3000**	**0.4000**	**0.4500**	0.4667	**0.4333**	**0.3833**	**0.4667**	0.2667
ID	29	30	31	32	33	34	35	36	37	38	39	40	41	42
Selected Backbone ID	19	19	19	19	19	19	19	19	19	19	19	19	19	19
Wo_Backbone	0.5167	0.6500	0.5500	0.5333	0.5333	0.4833	0.7167	0.4000	0.4167	0.6500	0.4000	0.5500	0.4667	0.3667
W_Backbone	**0.6667**	0.5667	**0.6167**	**0.5833**	**0.7333**	0.2333	**0.7667**	0.3500	**0.5000**	0.6333	**0.4500**	**0.5833**	**0.5667**	**0.4333**
ID	43	44	45	46	47	48	49	50	51	52	53	54	55	56
Selected Backbone ID	19	19	19	19	19	19	19	19	19	19	19	19	19	19
Wo_Backbone	0.3833	0.3333	0.2833	0.2667	0.5333	0.4833	0.1667	0.4333	0.5333	0.5000	0.3133	0.3167	0.3833	0.5500
W_Backbone	**0.5167**	**0.3833**	**0.3833**	0.2333	0.5333	**0.6333**	**0.3167**	0.4167	0.4333	0.5000	**0.3667**	**0.4500**	**0.5833**	0.3667

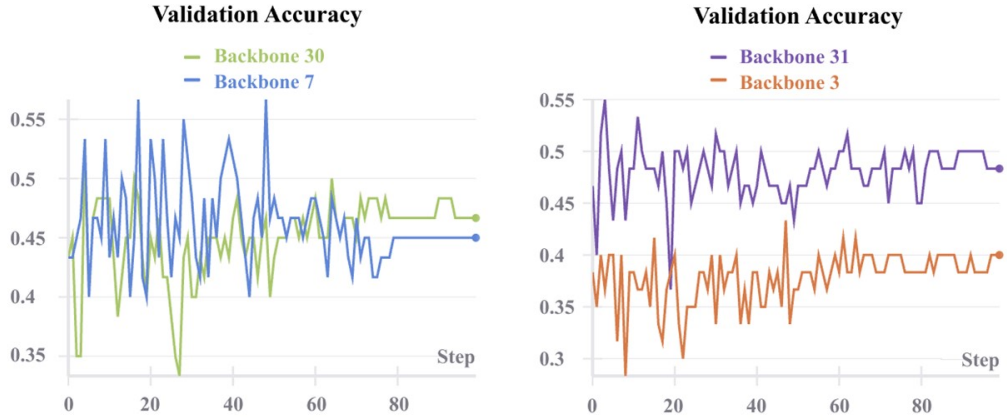

Fig. 2. Validation accuracy over training steps for target tasks 6 (left) and 2 (right). For both tasks, selecting a cross-modal backbone (30 for task 6, 31 for task 2) yields higher and more stable performance than using an intra-modal backbone (7 and 3, respectively), demonstrating the effectiveness of cross-modal selective transfer in *CrossBack*.

Oxford 102 Flowers Results. For tasks 1–6, which used the image modality, and tasks 7–12, which used the text modality, experiments were conducted. While Pick-a-back was originally proposed for image modality, the method showed strong performance in the text-based environment as well, with the model using `w_backbone` outperforming `wo_backbone` in most tasks. Specifically, task ID 11 showed an improvement of over 0.05 compared to`wo_backbone`, suggesting that the backbone selection strategy can be effectively applied to the text modality. Additionally, the use of the Perceiver IO model in the experiments further confirms that Pick-a-back is compatible with various architectures.

1) Intra-modal Transfer. Table 2 presents the results of intra-modal knowledge transfer using our proposed framework. These experiments were conducted in a setting where both the source and target tasks share the same modality (image or text), allowing us to evaluate the effectiveness of the selective backbone transfer strategy within a single modality. The method, referred to as *CrossBack*, demonstrates strong performance in both visual and textual domains.

We conducted experiments on 28 image-based tasks (Task IDs 1–28) and 28 text-based tasks (Task IDs 29–56). Among the image modality models, 18 out of 28 showed improved performance when using `w_backbone` compared to `wo_backbone`. For text-based models, 12 out of 28 showed gains. These results highlight that *CrossBack* performs effective backbone selection in an unimodal environment, especially for image modality tasks.

2) Cross-modal Transfer. Table 3 presents the results of cross-modal backbone selection using *CrossBack* on the CUB-200-2011 dataset. In this setting, a model is allowed to choose a backbone from a different modality, enabling evaluation of distribution-level compatibility across modalities.

We carried out experiments involving 28 tasks based on images (IDs 1–28) and 28 tasks based on text (IDs 29–56), utilizing selected backbone architectures that ap-

ply to either the image or text modality. Among the image modality models, 20 tasks achieved higher accuracy when a text-based backbone was selected, compared to using another image-based backbone. Conversely, among the text modality models, 18 tasks showed improved performance when selecting an image-based backbone. These results suggest that our strategy effectively performs backbone selection even under modality mismatches, guided by distributional similarity (DDV) rather than strict modality alignment. The performance gains in both directions indicate the modality-agnostic transferability and flexibility of *CrossBack*.

C. Discussion

The experimental results presented above provide valuable insights into the mechanisms and potential of *CrossBack*. Firstly, the selective transfer mechanism proved robust across modalities, indicating that distributional similarity—rather than modality alignment—is a reliable guide for knowledge reuse. This supports the core design principle of leveraging Distributional Distance Vectors (DDV) to assess task-level compatibility, instead of relying on handcrafted heuristics. Interestingly, while image-based tasks consistently benefited from selective backbone reuse, the improvements were less uniform across text-based tasks. This discrepancy may stem from the variability in representational quality across backbones trained on different modalities, particularly since text-based representations are often more sensitive to architectural choices and pretraining dynamics. Further investigation into modality-specific tuning could help mitigate these inconsistencies. Another significant insight is the strength of the Perceiver IO backbone in facilitating seamless interaction across diverse input formats. Its ability to process varying input sizes and structures enabled the unified architecture to be effective under heterogeneous conditions.

Despite these strengths, some limitations remain. For example, while our framework supports backbone selection across modalities, it currently does not account for potential degradation when the selected backbone introduces noise or incompatible features. Additionally, the scalability of the selection process in real-time or large-scale decentralized environments has yet to be thoroughly explored.

Even so, *CrossBack* offers two additional advantages critical for real-world deployment: (1) faster backbone retrieval via precomputed compatibility metrics, and (2) enhanced privacy through the transfer of only projection layers and intermediate features—never raw data or full weights. These properties further solidify *CrossBack*'s suitability for edge learning environments with strict latency and security constraints.

Future work may expand the method to include additional modalities such as audio and video or adapt it for sequence-to-sequence and generative tasks. Investigating adaptive or learnable backbone selection strategies, as opposed to relying solely on similarity measures, could further enhance robustness. Lastly, exploring collaborative backbone training in federated or self-supervised learning settings may unlock new possibilities for decentralized continual learning.

In summary, the discussion of results reveals that *CrossBack* offers a promising direction for enabling flexible and efficient knowledge reuse across diverse edge environments. With further refinement, it has the potential to become a foundational mechanism in real-world multimodal AI systems.

VII. Conclusion

CrossBack is a new framework for selective knowledge transfer on the edge, where devices process data of heterogeneous modalities, resolutions, and sequence lengths. Building on Pick-a-back, it assesses cross-modal compatibility, selects the peer whose decision patterns best align, and selectively reuses that backbone, thereby overcoming both input-size and modality mismatches.

We address a fundamental question: How can models with divergent input sizes and modalities effectively exploit external knowledge? Extensive experiments demonstrate that even models trained on orthogonal modalities (e.g., images and text) can exchange knowledge through *CrossBack*, simultaneously boosting accuracy and reducing computation. These findings move edge AI beyond single-modality assumptions and pave the way for continual, autonomous collaboration in decentralized environments, with the proposed *CrossBack* framework emerging as a core enabler for real-time multimodal applications.

References

[1] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26, 2013.

[2] Ching-Yi Hung, Cheng-Hao Tu, Cheng-En Wu, Chien-Hung Chen, Yi-Ming Chan, and Chu-Song Chen. Compacting, picking and growing for unforgetting continual learning. *Advances in neural information processing systems*, 32, 2019.

[3] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021.

[4] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[5] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

[6] Yuanchun Li, Ziqi Zhang, Bingyan Liu, Ziyue Yang, and Yunxin Liu. Modeldiff: Testing-based dnn similarity comparison for model reuse detection. In *Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis*, pages 139–151, 2021.

[7] Wenzhuo Liu, Fei Zhu, Longhui Wei, and Qi Tian. C-clip: Multimodal continual learning for vision-language model. In *The Thirteenth International Conference on Learning Representations*, 2025.

[8] Arun Mallya, Dillon Davis, and Svetlana Lazebnik. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In *Proceedings of the European conference on computer vision (ECCV)*, pages 67–82, 2018.

[9] Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7765–7773, 2018.

[10] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[11] Yasuhide Mori, Hironobu Takahashi, and Ryu ichi Oka. Image-to-word transformation based on dividing and vector quantizing images with words. 1999.

[12] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.

[13] Ariadna Quattoni, Michael Collins, and Trevor Darrell. Learning visual representations using images with captions. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[15] Nitish Srivastava and Russ R Salakhutdinov. Multimodal learning with deep boltzmann machines. *Advances in neural information processing systems*, 25, 2012.

[16] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[17] JinYi Yoon and HyungJune Lee. Pick-a-back: Selective device-to-device knowledge transfer in federated continual learning. In *European Conference on Computer Vision*, pages 165–182. Springer, 2024.

SUMMARY OF THIS PAPER

A. Problem Setup

Multi-modal AI offers robustness and accuracy by integrating data from vision, text, audio, and sensors. However, deploying such models on edge devices is challenging due to constraints on computation, memory, and security, as well as heterogeneity in input shape and modality. These factors render uniform model deployment and centralized updates impractical. Selective knowledge transfer addresses this by allowing each device to adopt only the external knowledge most compatible with its local context. Prior work, such as *Pick-a-back*, enables decentralized backbone reuse based on decision similarity, but assumes homogeneous input.

B. Novelty

We introduce *CrossBack*, the selective knowledge-transfer framework that jointly handles input-scale and modality heterogeneity in decentralized edge environments. Building on Pick-a-back, *CrossBack* leverages internal cross-attention key/value projections to compute DDV-cosine and DDV-euclidean distances, enabling each device to identify and reuse the most compatible peer backbone across both vision and language modalities and across varying input resolutions. Our fully decentralized design obviates any central server and employs dynamic piggymasks to manage capacity and prevent forgetting.

C. Algorithms

Require: $L, \{M_0^{(l)}, T^{(l)}, \theta^{(l)}, CL, \text{piggyMask}\}_{l=1}^{L}$
for $l = 1, \ldots, L$ **do**
  for $i = 1, \ldots, |T^{(l)}|$ **do**
    $M_{\text{loc}} \leftarrow CL.\text{train}(M_{i-1}^{(l)}, T_i^{(l)}, \theta_i^{(l)})$ ▷ local learning
    for $k = 1, \ldots, L$ **do**
      $\text{sim}[k] \leftarrow \text{combine}\Big(cos\_ddv\big(DDV(\text{attnRep}(M_{\text{loc}}, \text{extractKV}(M_{i-1}^{(k)}))), DDV(\text{attnRep}(M_{i-1}^{(k)}, \text{extractKV}(M_{i-1}^{(k)})))\big), euc\_ddv(\ldots)\Big)$ ▷ attention
    end for
    $k^* \leftarrow \arg\max_k \text{sim}[k]$
    $\hat{M} \leftarrow \text{init\_piggyMask}(M_{\text{loc}}, \text{extractKV}(M_{i-1}^{(k^*)}))$ ▷ fuse KV
    $M_i^{(l)} \leftarrow CL.\text{train\_with\_mask}(\hat{M}, T_i^{(l)}, \theta_i^{(l)})$ ▷ masked adaptation
  end for
end for
return $\{M_{|T^{(l)}|}^{(l)}\}_{l=1}^{L}$

Notation. L is the number of devices; $M_0^{(l)}$ initial Perceiver IO on device l; $T^{(l)}$ its task sequence; $\theta^{(l)}$ task hyperparameters; CL the continual-learning strategy; init_piggyMask injects a piggyMask over extracted KV layers; $\text{extractKV}(M)$ extracts the key–value pair from M; $\text{attnRep}(M, kv)$ returns the cross-attention output representation of model M given key–value input kv; cos_ddv, euc_ddv compute cosine- and Euclidean-based distributional distances of DDVs.

D. Experiments

We evaluated *CrossBack* in continual federated learning with decentralized edge devices with modality heterogeneity. Devices sequentially received image- or text-based tasks without centralized data or the parameters of other models. Experiments on CIFAR-100, CUB-200-2011, and Oxford 102 Flowers used modality-specific Perceiver IO models. Our pipeline included five stages: baseline training, pruning, backbone selection based on distributional similarity(DDV), backbone transfer, and final evaluation. We compared three settings: no backbone, random backbone, and *CrossBack*-selected backbone. Results show *CrossBack* consistently outperforms both baseline and random transfer in intra-modal and cross-modal tasks. It remains robust under modality mismatches, demonstrating the effectiveness of DDV. Additional ablations on CIFAR-100 and Oxford 102 Flowers confirmed compatibility with single-modal settings and prior methods like *Pick-a-back*.