CC-MBS: A Missing-Modality-Robust Multimodal Sample Selection Strategy for UAV Swarms

Yuntao Xu *, Bing Chen *, Feng Hu, Yue Cai, Zhuqing Xu

Posted Date: 3 June 2026

doi: 10.20944/preprints202606.0287.v1

Keywords: UAV swarm; edge intelligence; multimodal learning; sample selection

 doi:10.20944/preprints202606.0287.v1

Article

CC-MBS: A Missing-Modality-Robust Multimodal Sample Selection Strategy for UAV Swarms

Yuntao Xu *, Bing Chen *, Feng Hu, Yue Cai and Zhuqing Xu

College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China, 211106

* Correspondence: yuntaoxu@nuaa.edu.cn (Y.X.); cb_china@nuaa.edu.cn (B.C.)

Highlightsp

What are the main findings?

- A cross-UAV neighborhood collaborative compensation mechanism is proposed to enhance modality reliability by leveraging modality confidence exchange across distributed nodes, enabling effective information recovery under incomplete modalities.
- A modality-aware sample selection strategy is developed based on compensated modality confidence, allowing more reliable identification and retention of high-value samples under limited memory.

What are the implications of the main findings?

- The collaborative compensation mechanism demonstrates that multimodal robustness in UAV swarms can be significantly improved without high-dimensional feature or parameter sharing, by relying on lightweight confidence-level interaction across nodes.
- The sample selection strategy shows that incorporating modality quality and cross-node information leads to more stable and efficient data utilization, especially under modality missing and high pruning conditions.

Abstract

In resource-constrained UAV swarm systems, multimodal sensory data are often affected by complex environmental factors, resulting in modality missing, signal degradation, and asynchrony, which significantly reduce the reliability of multimodal learning and incremental model updates. To address this issue, we propose a Compensatory Collaboration Modality-Balanced Sample Selection framework (CC-MBS), which improves robustness through modality quality modeling and cross-UAV collaborative compensation. Specifically, a modality confidence vector is introduced to quantify modality reliability from missing rate, degradation, and asynchrony. A lightweight collaboration mechanism is designed to exchange low-dimensional confidence information instead of high-dimensional features or model parameters. Based on the compensated confidence, a modality-aware sample selection strategy is further developed to prioritize high-value samples under limited memory. Experimental results on multi-UAV scenarios show that CC-MBS outperforms representation-based methods such as ShaSpec and its parameter aggregation variants (AVG, PFM, POW) in modality compensation under missing conditions. In addition, it achieves stronger robustness than MBS and training-dynamics-based methods such as EL2N and GraNd in sample selection. These results demonstrate that CC-MBS effectively improves robustness and data efficiency for multimodal incremental learning under incomplete modalities.

Keywords: UAV swarm; edge intelligence; multimodal learning; sample selection

 doi:10.20944/preprints202606.0287.v1

1. Introduction

Unmanned aerial vehicle (UAV) swarms have garnered considerable attention due to their capability to perform distributed perception and decision-making in complex tasks such as disaster response, environmental monitoring, and intelligent surveillance. In these applications, multiple UAVs typically need to collaboratively acquire multimodal data under constrained computational, storage, and communication resources, in order to construct a comprehensive understanding of the environment.

However, in such resource-limited UAV swarm systems, multimodal data are often affected by occlusions, wind noise, illumination changes, and communication fluctuations, resulting in modality missing, signal degradation, and cross-modal asynchrony, which can significantly impair the overall perception quality and decision-making reliability of the swarm. Moreover, during long-term operation, the data received by UAV swarms exhibit streaming, non-stationary, and continuously evolving distributions, leading to both concept drift and modality drift, which render traditional offline learning methods based on static datasets inadequate for maintaining stable performance. Consequently, incremental learning mechanisms capable of continuously integrating new data and dynamically updating model parameters have become essential for enabling reliable perception and decision-making in UAV swarms.

Due to the limited computational and storage capacity of UAVs, it is infeasible to retain all historical data for model updates, necessitating the selective retention of key samples with high learning value from streaming data. Existing sample selection strategies often assume complete and reliable multimodal data, neglecting the widespread issues of modality incompleteness, degradation, and asynchrony in real-world UAV deployments. Such assumptions may introduce noisy or misleading samples, thereby reducing learning effectiveness and exacerbating catastrophic forgetting. Furthermore, compared to single-agent systems, sample selection in UAV swarms exhibits more complex system-level characteristics: variations in environmental conditions, sensor states, and communication quality across UAVs lead to heterogeneous data quality, making global optimal sample selection more challenging. Importantly, the degradation of a modality in one UAV can often be partially mitigated by information from neighboring UAVs; however, most existing methods do not explicitly model these cross-agent collaborative relationships. As a result, when UAVs encounter different environmental disturbances or sensor limitations, model updates fail to fully leverage swarm-level collaboration, leading to suboptimal performance. Therefore, accurately assessing sample learning value while exploiting collaborative advantages under modality incompleteness and dynamically varying data quality remains a critical and underexplored challenge.

To address these challenges, we propose a **Cooperative and Compensatory Modality-Balanced Sample Selection (CC-MBS)** framework, which enhances the robustness of UAV swarms under modality missingness through a cooperative compensation mechanism. First, we introduce a **modality confidence vector** to quantify the reliability of each modality at a given time, jointly considering modality missing rates, signal degradation levels, and cross-modal instability. This design enables each UAV to assess its modality status in real time and identify potential modality failures.

Second, we develop a lightweight cooperative compensation mechanism that allows UAVs to share low-dimensional modality confidence vectors under controlled communication overhead. By leveraging cross-agent perceptual diversity, UAVs can compensate for degraded modalities through collaboration. This process enables the swarm to maintain stable perception capabilities even when individual UAVs suffer from modality degradation, thereby supporting accurate multimodal learning and decision-making in complex environments.

Building upon this, we propose the **CC-MBS sample selection strategy**, which aims to retain the most informative samples for model improvement under resource-constrained incremental learning. Consistent with our previous Modality-Balanced Sample Selection (MBS) strategy [1], CC-MBS evaluates sample value based on modality balance, but further incorporates cooperative compensation information from other UAVs to more accurately estimate the actual contribution of

 doi:10.20944/preprints202606.0287.v1

each sample to model updates. Through this collaborative evaluation mechanism, CC-MBS ensures that only highly informative and reliable samples are retained in the replay buffer, effectively mitigating catastrophic forgetting and enabling continuous adaptation to dynamic environments. Moreover, the introduction of cooperative compensation allows UAV swarms to better exploit cross-agent perceptual diversity, maintaining stable learning performance even when individual sensing quality fluctuates over time. This design makes CC-MBS a natural extension of our previous MBS [1] framework, providing a more robust solution for incremental learning in multi-UAV systems under complex multimodal scenarios.

The main contributions of this work are summarized as follows:

- We propose a cooperative compensatory multimodal sample selection framework (CC-MBS) for UAV swarm perception under modality missingness and dynamic environments, which unifies modality quality modeling, cross-UAV collaboration, and sample selection into a single framework, significantly improving robustness and adaptability;
- We develop a unified modality confidence modeling method that characterizes modality reliability from multiple aspects, including modality missingness, signal degradation, and cross-modal asynchrony, providing an interpretable and computable basis for both cooperative compensation and sample valuation;
- We design a cooperative compensatory modality-balanced sample selection strategy, which effectively alleviates performance degradation and catastrophic forgetting under incomplete modalities. Experiments conducted on small-scale UAV swarms (2–5 agents) demonstrate that the proposed method achieves strong stability and effectiveness across various modality missing rates.

2. Related Work

2.1. Cooperative Learning in UAV Swarms

With the increasing deployment of UAV swarms in environmental perception and mission execution, enabling efficient cooperative learning under limited computational and communication resources has become an important research topic. Existing work mainly focuses on two directions: distributed learning and cooperative perception.

In distributed learning, federated learning has been widely adopted in multi-UAV systems to achieve data privacy preservation and communication efficiency. For example, prior studies apply federated learning to UAV networks for distributed model training and knowledge sharing [2], and further extend it to hierarchical federated learning frameworks to accommodate complex network topologies and heterogeneous nodes [3]. In addition, collaborative learning approaches incorporating blockchain and incentive mechanisms have been proposed to improve reliability and cooperation efficiency in UAV systems [4].

On the other hand, in cooperative perception and decision-making, researchers exploit the spatial diversity of UAVs to enhance overall perception through multi-view information fusion. Recent studies show that UAVs can improve modeling accuracy of target trajectories and scene structures by sharing local observations [5,6], and achieve coordinated optimization in tasks such as target search and tracking using multi-agent reinforcement learning [7]. These works demonstrate that cross-node information complementarity can effectively alleviate the limitations of individual perception.

However, existing UAV collaborative learning methods primarily focus on model parameter sharing, task allocation, or feature-level fusion, often assuming that information from different nodes is homogeneous or equivalent, and lack explicit modeling of differences in multimodal data quality. In particular, in multimodal perception scenarios, the observation quality of different UAVs across various modalities often exhibits significant variability, making the question of how to effectively leverage cross-node information for compensation a critical yet underexplored issue.

 doi:10.20944/preprints202606.0287.v1

2.2. Modality Missingness and Degradation in Multimodal Perception

In multimodal perception systems, most existing methods assume that all modalities are complete and synchronously available during both training and inference. However, this assumption is often violated in real-world applications, particularly in dynamic environments such as UAV swarms, where occlusions, sensor failures, environmental noise, and communication instability frequently lead to modality missingness and incomplete information.

Early studies have demonstrated that multimodal models heavily depend on modality completeness. For example, HeMIS [8] shows that modality missingness at test time can significantly degrade model performance. Subsequent works extend this observation to more realistic scenarios. SMIL [9] and MMIN [10], from the perspectives of severe missingness and uncertain missingness respectively, demonstrate that model performance deteriorates substantially under random multimodal missing patterns, highlighting modality incompleteness as a fundamental challenge in multimodal learning.

Beyond complete modality missingness, recent studies have increasingly focused on modality degradation and cross-modal asynchrony. GMC [11] indicates that noise and geometric inconsistency across modalities in feature space can weaken cross-modal alignment. Further studies suggest that real-world multimodal systems often simultaneously suffer from modality missingness, degradation, and temporal asynchrony [12]. At the application level, works such as M3AE [13] and MissModal [14] validate the persistent negative impact of modality incompleteness on model performance, while more recent research [15] emphasizes that the robustness bottleneck of multimodal systems arises from the compounded effects of multiple forms of incompleteness.

In summary, modality missingness, degradation, and asynchrony are not isolated issues but jointly constitute the core challenges in real-world multimodal perception. This observation provides strong motivation for the development of robust multimodal learning methods.

2.3. Existing Methods for Robustness to Modality Missingness

To address modality missingness and incompleteness in multimodal data, existing approaches mainly fall into three categories: representation learning, modality completion, and robust fusion.

One line of work focuses on learning modality-invariant shared representations to reduce reliance on specific modalities. For instance, HeMIS [8] aggregates statistical features across modalities to enable robust segmentation under missing modalities. GMC [11] enforces geometric consistency across modalities through contrastive learning, allowing stable representations even when some modalities are missing. Additionally, some methods improve adaptability to varying modality combinations via shared–specific feature decomposition. For example, ShaSpec [16] decomposes multimodal representations into shared and modality-specific components, preserving useful information under modality missingness.

Another line of research explicitly models and reconstructs missing modalities. MMIN [10] introduces a “modality imagination” mechanism to reconstruct latent representations of missing modalities from available ones, while SMIL [9] leverages meta-learning to improve generalization across diverse missing patterns. With the rise of pretrained multimodal models, recent works further explore robustness via prompting and lightweight adaptation. For instance, multimodal prompting methods [17] and their extensions [18] enable models to adapt to different modality combinations without full retraining. Moreover, self-supervised and distillation-based strategies have been widely adopted to unify multiple missing patterns. M3AE [13] learns cross-modal reconstruction relationships via masked autoencoding, while UMDF [19] employs self-distillation to handle various modality missing scenarios within a single model.

Although the aforementioned methods have achieved significant progress in single-agent multimodal learning scenarios, most studies still primarily focus on “within-sample” modality missing, paying limited attention to the effects of dynamic changes in modality quality and cross-device heterogeneity. Moreover, these methods generally assume that all modalities originate from the same data source, lacking explicit modeling of cross-node collaborative compensation in

 doi:10.20944/preprints202606.0287.v1

distributed systems. More importantly, existing approaches often concentrate on improvements at the model architecture or representation learning level, while giving less consideration to how high-value samples can be selected based on modality quality and collaborative information under resource-constrained, streaming data conditions. Therefore, in UAV swarm scenarios, how to jointly model modality quality, cross-node collaborative compensation, and sample selection to achieve more robust incremental learning remains an underexplored challenge.

3. Basic Model

To formally characterize the streaming data processing paradigm in UAV swarm scenarios and provide a unified theoretical foundation for subsequent sample selection strategies, we first establish a multimodal incremental learning framework for UAV swarms. Specifically, we define the multimodal sample representation, feature encoding, and fusion process, and adopt the optimization objective from prior work as the incremental learning objective.

Furthermore, to model the practical challenges commonly observed in real-world perception—namely modality missingness, modality degradation, and cross-modal asynchrony—we characterize modality status from three aspects: missing rate, degradation level, and asynchrony rate. Based on these factors, we construct a modality confidence vector, which serves as the foundation for subsequent cooperative compensation and sample selection strategies.

3.1. Multimodal Incremental Learning in UAV Swarms

Consider a UAV swarm consisting of Nagents, denoted as $\mathcal{U} = \{u_i\}_{i=1}^{N}$. During task execution, UAVs continuously collect multimodal sensory data and adopt a "lightweight local preprocessing + edge-side incremental learning" paradigm for model adaptation. Specifically, each UAV performs lightweight operations such as modality quality estimation, feature extraction, and modality confidence computation, while the edge node conducts online or near-online model updates over streaming data under resource constraints. Meanwhile, a capacity-limited replay buffer is maintained to mitigate catastrophic forgetting.

This architecture aligns with the practical characteristics of UAV swarms, including low latency requirements, unstable communication, and limited onboard computation. It also provides a unified system context for subsequent sample selection strategies, whose goal is to identify the most valuable samples under constrained computation and bandwidth. The overall framework is illustrated in Figure 1.

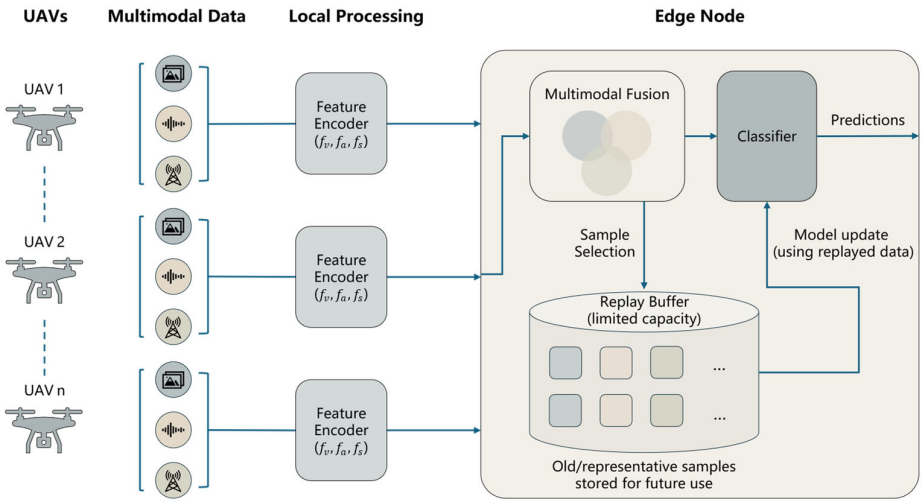


Figure 1. Multimodal incremental learning framework in a UAV swarm system.

 doi:10.20944/preprints202606.0287.v1

Let $\mathcal{M}$ denote the modality set. In typical UAV perception tasks, $\mathcal{M} = \{v, a, s\}$(vision, audio, and sensor modalities). In certain experimental settings, this can be simplified to bimodal scenarios $\mathcal{M} = \{v, a\}$, where industrial-grade sensors are assumed to be relatively stable, while visual and acoustic data are more susceptible to environmental disturbances.

At each discrete time step t, UAV u_i generates a labeled (or pseudo-labeled) multimodal sample:

$$x_{i,t} = \left(\{x_{i,t}^m\}_{m\in\mathcal{M}}, y_{i,t}\right), \tag{1}$$

where $x_{i,t}^m$ denotes the raw observation of modality m(e.g., image frames, spectrograms, or sensor vectors), and $y_{i,t} \in \mathcal{Y}_t$ is the corresponding label at time t. The data collected by the swarm at time t forms a swarm batch:

$$\mathcal{B}_t = \{x_{i,t}\}_{i=1}^N, \tag{2}$$

which constitutes a streaming multimodal data sequence $\{\mathcal{B}_t\}_{t=1}^T$. Due to resource constraints, the edge node cannot store the entire historical data stream, necessitating selective compression and retention of samples within each time window.

For each modality $m \in \mathcal{M}$, we define a modality encoder $E_m(\cdot)$ that maps raw observations into a shared feature space:

$$z_{i,t}^m = E_m(x_{i,t}^m) \in \mathbb{R}^d \tag{3}$$

The multimodal representations are then fused via a fusion function $f_\theta(\cdot)$:

$$z_{i,t} = f_\theta(\{z_{i,t}^m\}_{m\in\mathcal{M}}) \in \mathbb{R}^d \tag{4}$$

A classification head $g(\cdot)$ produces the predictive distribution:

$$P_{i,t} = \mathrm{softmax}\left(g(z_{i,t})\right), \tag{5}$$

where f_θ can be instantiated as any differentiable fusion operator, such as concatenation, summation, gated fusion, or attention-based fusion. This abstraction decouples the proposed sample selection strategy from specific network architectures, allowing it to be integrated as an external module.

To alleviate catastrophic forgetting in incremental learning, the edge node maintains a replay buffer $\mathcal{R}_t$with limited capacity, consisting of selected historical samples. Upon receiving the swarm batch $\mathcal{B}_t$, model parameters are updated by minimizing the following objective:

$$\mathcal{L}_t = \frac{1}{|\mathcal{B}_t|}\sum_{(x,y)\in\mathcal{B}_t} \mathcal{L}_{CE}\left(P(x), y\right) + \lambda \cdot \frac{1}{|\mathcal{R}_t|}\sum_{x\in\mathcal{R}_t} \mathcal{L}_{KD}\left(P(x), P_{\mathrm{pre}}(x)\right), \tag{6}$$

where $\mathcal{L}_{CE}$ denotes the cross-entropy loss, $\mathcal{L}_{KD}$ represents the knowledge distillation loss, $P_{\mathrm{pre}}(x)$ is the prediction of the previous model, and λ is a balancing coefficient.

3.2. Modality Missingness and Degradation Modeling

3.2.1. Modality Missingness Rate

In UAV swarm missions, multimodal observations may suffer from intermittent missingness due to occlusions, strong illumination, wind noise, sensor failures, and communication interruptions.

For UAV u_i, we define a modality availability indicator for modality $m \in \mathcal{M}$ at sampling index k within time window t(containing Ksamples) as:

$$\delta_{i,t,k}^m \in \{0,1\}, k = 1,2,\dots,K, \tag{7}$$

where $\delta_{i,t,k}^m = 1$ indicates that modality m is successfully observed at sampling point k, and $\delta_{i,t,k}^m = 0$ otherwise.

Based on this, the **modality missingness rate** is defined as:

 doi:10.20944/preprints202606.0287.v1

$$r_{i,t}^{m} = 1 - \frac{1}{K}\sum_{k=1}^{K} \delta_{i,t,k}^{m}, r_{i,t}^{m} \in [0,1], \tag{8}$$

a higher value of $r_{i,t}^{m}$ indicates more severe modality missingness within the time window. Specifically, $r_{i,t}^{m} \to 1$ implies that modality m is nearly unavailable, while $r_{i,t}^{m} \to 0$ indicates stable modality acquisition.

3.2.2. Modality Degradation Level

Beyond complete modality missingness, a more common scenario in real-world systems is modality degradation, where observations are available but of reduced quality. Typical examples include motion blur or overexposure in vision, reduced signal-to-noise ratio in audio, and sensor drift or packet loss in environmental sensing.

To provide a unified characterization of degradation across different modalities, we introduce a modality quality assessment function $Q_m(\cdot)$, which evaluates the raw quality score of each observation:

$$q_{i,t,k}^{m,\text{raw}} = Q_m(x_{i,t,k}^{m}), \tag{9}$$

can be instantiated using modality-specific quality metrics, such as blur estimation for vision, signal-to-noise ratio for audio, or packet integrity for sensor data.

To eliminate scale differences across modalities, we normalize the quality score to the range $[0, 1]$:

$$q_{i,t,k}^{m} = \frac{q_{i,t,k}^{m,\text{raw}} - q_{m}^{\min}}{q_{m}^{\max} - q_{m}^{\min}}, q_{i,t,k}^{m} \in [0,1], \tag{10}$$

where $q_{m}^{\min}$ and $q_{m}^{\max}$ are modality-specific normalization constants, estimated either from training data statistics or via online sliding-window estimation.

We then define the **modality degradation level** at the time-window level as:

$$d_{i,t}^{m} = 1 - \frac{\sum_{k=1}^{K} \delta_{i,t,k}^{m} q_{i,t,k}^{m}}{\sum_{k=1}^{K} \delta_{i,t,k}^{m} + \varepsilon}, d_{i,t}^{m} \in [0,1], \tag{11}$$

where ε is a small constant to avoid division by zero. This formulation ensures that the degradation level increases when either: the quality of available observations decreases, or the number of valid observations becomes limited.

3.2.3. Modality Asynchrony Rate

In multimodal sensing systems, different modalities often operate with heterogeneous sampling rates and transmission delays. As a result, observations corresponding to the same semantic event may exhibit temporal misalignment across modalities, which can disrupt cross-modal correspondence and degrade fusion performance.

Let $\tau_{i,t,k}^{m}$ denote the timestamp of modality $m \in \mathcal{M}$ for UAV u_i at the k-th sampling point within time window t. Given a reference modality m_{ref}(e.g., vision), we define an asynchrony indicator at each sampling point as:

$$z_{i,t,k}^{m} = \mathbb{I}\left(\left| \tau_{i,t,k}^{m} - \tau_{i,t,k}^{m_{\text{ref}}} \right| > \epsilon \right), \tag{12}$$

where $\mathbb{I}(\cdot)$ is the indicator function, and ϵ is a predefined synchronization tolerance threshold. ϵ can be determined based on sensor synchronization precision or empirically tuned.

Based on this, the **modality asynchrony rate** at the time-window level is defined as:

$$a_{i,t}^{m} = \frac{1}{K}\sum_{k=1}^{K} z_{i,t,k}^{m}, a_{i,t}^{m} \in [0,1] \tag{13}$$

 doi:10.20944/preprints202606.0287.v1

A higher value of $a_{i,t}^m$ indicates more severe temporal misalignment between modality $m$and the reference modality. Specifically, $a_{i,t}^m \to 1$ implies that modality $m$is largely unsynchronized and difficult to align during fusion, while $a_{i,t}^m \to 0$ indicates stable cross-modal temporal alignment.

3.2.4. Modality Confidence Vector

Based on the modality missingness rate $r_{i,t}^m$, degradation level $d_{i,t}^m$, and asynchrony rate $a_{i,t}^m$, we define the modality confidence for modality $m \in \mathcal{M}$ of UAV u_i within time window t as:

$$c_{i,t}^m = (1 - r_{i,t}^m)\,(1 - d_{i,t}^m)\,(1 - a_{i,t}^m), c_{i,t}^m \in [0,1] \tag{14}$$

This formulation provides an intuitive and interpretable measure of modality reliability: degradation in any of the three factors—availability, quality, or temporal alignment—will reduce the overall confidence. In particular, when a modality is highly missing, severely degraded, or strongly asynchronous, its confidence approaches zero.

Stacking the confidence values of all modalities yields the **modality confidence vector**:

$$\mathrm{c}_{i,t} = \left[c_{i,t}^m\right]_{m\in\mathcal{M}} \in [0,1]^{|\mathcal{M}|} \tag{15}$$

For subsequent fusion weighting and sample selection, the confidence vector can be normalized into a **modality weight vector**:

$$\omega_{i,t} = \frac{\mathrm{c}_{i,t}}{\sum_{m\in\mathcal{M}} c_{i,t}^m + \varepsilon}, \tag{16}$$

where ε is a small constant for numerical stability.

Illustrative Example:

Consider a time window where the visual modality has $r^v = 0.1$, $d^v = 0.3$, $a^v = 0.2$, yielding: $c^v = (1-0.1)(1-0.3)(1-0.2) = 0.504$; Similarly, for the audio modality with $r^a = 0.0$, $d^a = 0.5$, $a^a = 0.1$, we obtain: $c^a = (1-0)(1-0.5)(1-0.1) = 0.45$; Thus, the confidence vector is: $\mathbf{c} = [0.504, 0.45]$, and the normalized weight vector is approximately: $\omega \approx [0.528, 0.472]$.

The modality confidence vector serves as a key input to the subsequent **robust sample selection strategy under modality missingness**. When the quality of a modality deteriorates, its corresponding weight is automatically reduced, and the estimated sample value is accordingly adjusted. This mechanism effectively prevents noisy samples—caused by degraded or asynchronous modalities—from being excessively stored in the replay buffer, thereby improving the robustness of incremental learning.

4. Main Method

4.1. Multimodal Collaborative Compensation Mechanism Under UAV Swarms

During multimodal perception in UAV swarms, the reliability of the same modality can vary significantly across different UAVs due to environmental disturbances, viewpoint differences, and heterogeneous sensor conditions. To address this issue, we propose a multimodal collaborative compensation mechanism that integrates **neighborhood weighted aggregation** with a **gating mechanism**, enabling UAVs to selectively leverage neighboring information when local modality reliability degrades.

Consider a UAV swarm consisting of N agents, denoted as $\mathcal{U} = \{u_i\}_{i=1}^N$. At time window t, each UAV u_i computes its modality confidence vector: $\mathrm{c}_{i,t} = [c_{i,t}^m]_{m\in\mathcal{M}}, c_{i,t}^m \in [0,1]$, as defined in Eqs. (13)-(14).

To reflect realistic deployment constraints and avoid excessive global communication overhead, we adopt neighborhood-level collaboration rather than global aggregation. Specifically, UAVs can only collaborate with nearby agents that satisfy communication constraints. The neighborhood of UAV $u_i$is defined as:

 doi:10.20944/preprints202606.0287.v1

$$\mathcal{N}(i) = \{u_j \mid \mathrm{dist}(u_i, u_j) \le R\}, \tag{17}$$

where R is the communication radius, and $\mathrm{dist}(\cdot,\cdot)$ denotes spatial or communication distance.

For each modality $m \in \mathcal{M}$, we define collaborative weights based on modality confidence:

$$\omega_{i,j,t}^{m} = \frac{c_{j,t}^{m}}{\sum_{u_k \in \mathcal{N}(i)} c_{k,t}^{m} + \varepsilon}, u_j \in \mathcal{N}(i), \tag{18}$$

where ε is a small constant for numerical stability.

This formulation ensures that neighbors with higher modality confidence contribute more to the compensation process, while the normalization within the neighborhood reflects collective compensation rather than reliance on a single optimal node.

We further introduce a modality gating function to trigger collaboration only when necessary. Specifically:

$$g_{i,t}^{m} = \mathbb{I}(c_{i,t}^{m} < \eta_m), \tag{19}$$

where $\eta_m \in (0,1)$ is a modality-specific confidence threshold. If $c_{i,t}^{m} \ge \eta_m$, the local modality is considered reliable, and no compensation is applied; If $c_{i,t}^{m} < \eta_m$, the modality is deemed unreliable, and collaborative compensation is activated. This on-demand compensation mechanism avoids unnecessary communication and preserves efficiency.

Based on the above design, the compensated modality confidence is defined as:

$$\tilde{c}_{i,t}^{m} = c_{i,t}^{m} + g_{i,t}^{m} \cdot \lambda \sum_{u_j \in \mathcal{N}(i)} \omega_{i,j,t}^{m}\ c_{j,t}^{m}, \tag{20}$$

where $\lambda \in [0,1]$is a compensation strength coefficient. When $g_{i,t}^{m} = 0$, no compensation is applied, i.e., $\tilde{c}_{i,t}^{m} = c_{i,t}^{m}$; When $g_{i,t}^{m} = 1$, the compensated confidence incorporates a weighted aggregation of neighboring modality confidence.

This design follows an adaptive and demand-driven compensation principle, enabling UAVs to maintain stable modality reliability under dynamic and heterogeneous sensing conditions. Figure 2 illustrates the overall collaborative compensation process.

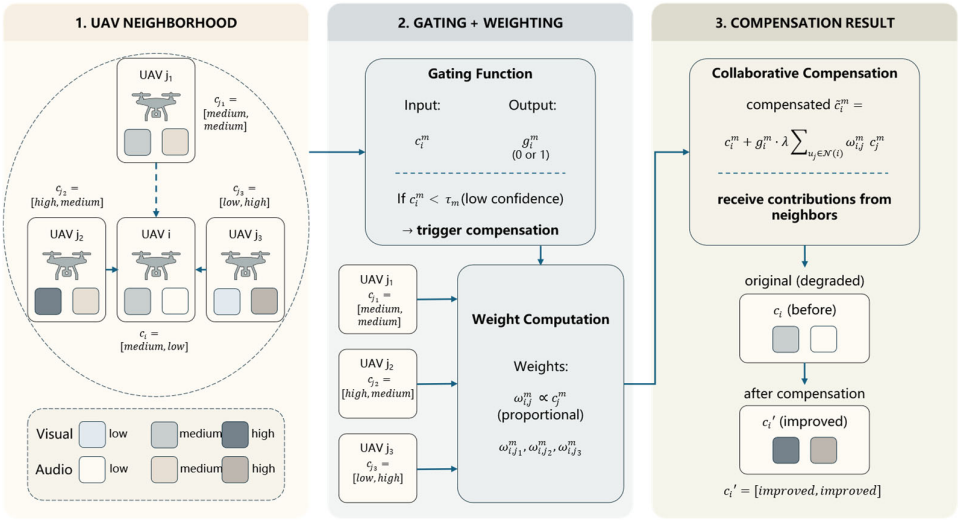


Figure 2. Multimodal collaborative compensation mechanism in UAV swarm.

After collaborative compensation, the modality confidence values are stacked to form the compensated modality confidence vector:

$$\tilde{c}_{i,t} = [\tilde{c}_{i,t}^{m}]_{m \in \mathcal{M}} \tag{21}$$

 doi:10.20944/preprints202606.0287.v1

This vector serves as a fundamental input to the subsequent CC-MBS sample selection strategy, where it is used to evaluate the overall reliability of each sample from a "local + collaborative compensation" perspective.

The complete collaborative compensation process is summarized in Algorithm 1.

Algorithm 1: Multimodal Collaborative Compensation Mechanism

Input:

- Local modality confidence $c_{i,t} = [c_{i,t}^m]_{m \in M}$;
- neighbor set $\mathcal{N}(i)$;
- thresholds η_m;
- compensation strength λ;
- constant ε.

Output:

- Compensated confidence vector $\tilde{c}_{i,t} = [\tilde{c}_{i,t}^m]_{m \in M}$.

Steps:

Initialize: $\tilde{c}_{i,t} \leftarrow c_{i,t}$

for each modality $m \in \mathcal{M}$ **do**

(a) Gating:

$g_{i,t}^m \leftarrow \mathbb{I}(c_{i,t}^m < \eta_m)$

(b) Compute normalization term:

$Z \leftarrow \sum_{u_k \in \mathcal{N}(i)} c_{k,t}^m + \varepsilon$

(c) Neighborhood aggregation:

$$\tilde{c}_{i,t}^m \leftarrow \sum_{u_j \in \mathcal{N}(i)} \frac{c_{j,t}^m}{Z} \cdot c_{j,t}^m$$

(d) Gated compensation update:

$\tilde{c}_{i,t}^m \leftarrow c_{i,t}^m + g_{i,t}^m \cdot \lambda \cdot \tilde{c}_{i,t}^m$

end for

return $\tilde{c}_{i,t}$

4.2. Compensatory Collaborative Modality-Balanced Sample Selection (CC-MBS)

Building upon the collaborative compensation mechanism introduced in Section 4.1, we propose a **Compensatory Collaborative Modality-Balanced Sample Selection (CC-MBS)** strategy for UAV swarms. This method extends our prior Modality-Balanced Sample Selection (MBS) framework [1] by explicitly incorporating both local modality confidence and neighborhood-level collaborative compensation into the sample valuation process. As a result, CC-MBS can reliably identify high-value samples even under modality missingness, degradation, and asynchrony.

Following the prior MBS framework, we maintain class prototypes in each modality space and employ a non-parametric prototype classifier to measure modality-wise discriminability.

For a sample generated by UAV u_i at time t: $x_{i,t} = (\{x_{i,t}^m\}_{m \in \mathcal{M}}, y_{i,t})$, we denote the encoded modality feature as: $z_{i,t}^m = E_m(x_{i,t}^m)$, and the class prototype for class y in modality m as c_y^m.

The modality-level prototype confidence is defined as:

$$s_{i,t}^m = \frac{\exp\left(-\mathrm{dist}(z_{i,t}^m, c_{y_{i,t}}^m)\right)}{\sum_{y' \in \mathcal{Y}_t} \exp\left(-\mathrm{dist}(z_{i,t}^m, c_{y'}^m)\right)}, m \in \mathcal{M}, \tag{22}$$

where $\mathrm{dist}(\cdot,\cdot)$ denotes a distance metric such as Euclidean or cosine distance.

 doi:10.20944/preprints202606.0287.v1

To quantify modality balance, the original MBS is defined based on the dispersion of modality confidences:

$$\Delta_{i,t} = \max_{m \in \mathcal{M}} s_{i,t}^{m} - \min_{m \in \mathcal{M}} s_{i,t}^{m}, \tag{23}$$

$$\mathrm{MBS}_{i,t} = 1 - \frac{\Delta_{i,t}}{\sum_{m \in \mathcal{M}} s_{i,t}^{m} + \varepsilon}, \mathrm{MBS}_{i,t} \in [0,1] \tag{24}$$

Intuitively, samples with balanced modality contributions (i.e., similar $s_{i,t}^{m}$) yield higher MBS scores, while dominance by a single modality leads to lower scores.

However, the original MBS assumes complete and reliable modalities, which is unrealistic in UAV swarm scenarios. Under modality degradation, relying solely on $s_{i,t}^{m}$ may lead to misjudgment. For example, a degraded visual modality may still produce high confidence due to spurious features, resulting in incorrect modality dominance.

To address this issue, we incorporate the **collaboratively compensated modality confidence** $\tilde{c}_{i,t}^{m}$ as an explicit reliability prior and define:

$$\hat{s}_{i,t}^{m} = \tilde{c}_{i,t}^{m} \cdot s_{i,t}^{m}, m \in \mathcal{M} \tag{25}$$

This formulation enables confidence-aware reweighting, where unreliable modalities are automatically suppressed, while reliable or collaboratively compensated modalities are emphasized.

Based on the reweighted modality confidences, we redefine the modality discrepancy:

$$\widehat{\Delta}_{i,t} = \max_{m \in \mathcal{M}} \hat{s}_{i,t}^{m} - \min_{m \in \mathcal{M}} \hat{s}_{i,t}^{m}, \tag{26}$$

and construct the **CC-MBS score**:

$$\text{CC-MBS}_{i,t} = 1 - \frac{\widehat{\Delta}_{i,t}}{\sum_{m \in \mathcal{M}} \hat{s}_{i,t}^{m} + \varepsilon} \tag{27}$$

Through this design, CC-MBS jointly captures: modality discriminability (via prototype confidence), modality reliability (via confidence modeling), and cross-agent compensation (via collaborative confidence).

At time window t, the edge node receives a batch of samples from the UAV swarm: $\mathcal{B}_t = \{x_{i,t}\}_{i=1}^{N}$, and maintains a replay buffer $\mathcal{R}_t$with capacity C.

To select the most valuable samples, we merge the current batch with the buffer and rank all samples by CC-MBS:

$$\mathcal{R}_{t+1} = \mathrm{TopK}(\mathcal{B}_t \cup \mathcal{R}_t,\ \text{CC-MBS},\ C), \tag{28}$$

where $\mathrm{TopK}(\cdot)$ selects the top-C samples with the highest CC-MBS scores.

The updated buffer $\mathcal{R}_{t+1}$ is then used in the knowledge distillation term of Eq. (6), ensuring that high-value historical samples are retained to mitigate catastrophic forgetting.

Overall, CC-MBS tightly integrates modality quality modeling, collaborative compensation, and sample selection, enabling robust incremental learning under modality incompleteness and resource constraints. The full procedure is summarized in Algorithm 2.

Algorithm 2: Compensatory Collaboration Modality-Balanced Sample Selection

Input:

- Batch data $\mathcal{B}_t = \{x_{i,t}\}_{i=1}^{N}$;
- Replay buffer $\mathcal{R}_t$ with capacity C;
- Modality encoders $E_m(\cdot)$;
- Class prototypes $\{c_y^m\}$;
- Compensated confidence vectors $\tilde{c}_{i,t} = [\tilde{c}_{i,t}^{m}]$;

 doi:10.20944/preprints202606.0287.v1

- Constant ε.

Output:

- Updated replay buffer $\mathcal{R}_{t+1}$.

Steps:

Initialize: $\mathcal{U}_t \leftarrow \mathcal{B}_t \cup \mathcal{R}_t$

for each sample $x_{i,t} \in \mathcal{U}_t$ **do**

(a) Modality feature encoding:

$z_{i,t}^m \leftarrow E_m(x_{i,t}^m), \forall m \in \mathcal{M}$

(b) Prototype-based modality confidence:

$$s_{i,t}^m \leftarrow \frac{\exp\left(-\mathrm{dist}(z_{i,t}^m, c_{y_{i,t}}^m)\right)}{\sum_{y' \in \mathcal{Y}_t} \exp\left(-\mathrm{dist}(z_{i,t}^m, c_{y'}^m)\right)}$$

(c) Confidence-aware reweighting:

$\hat{s}_{i,t}^m \leftarrow \tilde{c}_{i,t}^m \cdot s_{i,t}^m$

(d) Modality discrepancy computation:

$$\widehat{\Delta}_{i,t} \leftarrow \max_{m \in \mathcal{M}} \hat{s}_{i,t}^m - \min_{m \in \mathcal{M}} \hat{s}_{i,t}^m$$

(e) CC-MBS score:

$$\text{CC-MBS}_{i,t} \leftarrow 1 - \frac{\widehat{\Delta}_{i,t}}{\sum_{m \in \mathcal{M}} \hat{s}_{i,t}^m + \varepsilon}$$

end for

$\mathcal{R}_{t+1} \leftarrow \mathrm{TopK}(\mathcal{U}_t, \text{CC-MBS}, C)$

return $\mathcal{R}_{t+1}$

5. Evaluation

5.1. Experimental Setup

5.1.1. Hardware, System Setup, and Datasets

We evaluate our method on a prototype UAV swarm system following a device–edge collaborative architecture.

On the device side, we deploy multiple small rotary-wing UAVs as distributed sensing agents. Each UAV is equipped with an embedded computing platform (NVIDIA Jetson Orin) and multimodal sensing modules, including RGB cameras, microphone arrays, and basic environmental sensors. These devices perform lightweight local processing, such as modality quality estimation and feature extraction. Due to the limited computational capacity and energy constraints of UAV platforms, no heavy model training is conducted on-device.

On the edge side, we use a high-performance workstation as a centralized edge node for model training and incremental updates. The system is equipped with two NVIDIA RTX 4090 GPUs (24GB memory each), an AMD EPYC 7B12 processor, and 128GB RAM. The software environment is based on Ubuntu 24.04, with Python 3.9 and PyTorch 2.2, supported by CUDA 12.1 and cuDNN 8.9.

During operation, UAVs transmit preprocessed modality features or modality confidence information to the edge node via wireless communication (Wi-Fi). The edge node performs collaborative compensation and sample selection over streaming data, and updates the model using a capacity-limited replay buffer. This "edge training + on-device perception" architecture reflects realistic deployment constraints and provides a practical foundation for the proposed CC-MBS framework.

To ensure experimental controllability and reproducibility, we focus on bimodal (audio–visual) settings and simulate modality missingness and degradation at the feature level. Instead of real-time

 doi:10.20944/preprints202606.0287.v1

data collection, we adopt public datasets due to their diversity, clear annotations, and flexibility in controlling modality corruption.

We use two widely adopted multimodal datasets:

CREMA-D[20] is a multimodal dataset for emotion recognition, containing 7,442 audio–visual clips performed by 91 actors of diverse genders, ages, and ethnicities. Each clip expresses one of six basic emotions (anger, disgust, fear, happiness, neutral, sadness) plus surprise, under varying intensities and contexts. The dataset provides aligned audio and visual modalities, making it suitable for cross-modal perception and fusion tasks.

AVE[21] is a multimodal dataset for audio–visual event recognition, containing over 4,000 10-second clips collected from YouTube, covering 28 real-world event categories (e.g., dog barking, drum playing, crowd cheering). A key characteristic is the varying modality dependency across events—some rely more on audio (e.g., thunder), while others rely more on visual cues (e.g., instrument playing), making it well-suited for evaluating modality imbalance and robustness.

5.1.2. Experimental Scenarios

To systematically evaluate the proposed method, we construct multiple multimodal missingness and collaborative compensation scenarios within a unified simulation framework. Given the streaming nature of UAV data, we organize data into time windows t, each divided into 10 consecutive blocks to simulate short-term continuous observations. This design captures temporal dynamics and allows controlled injection of modality corruption over time.

We assume that UAVs operate within the same task region and observe a common target or scene (e.g., the same event or environment). However, due to viewpoint differences, environmental disturbances, and sensor variability, the quality of observations differs across UAVs and modalities. This setting reflects practical UAV applications such as target monitoring and event perception, while ensuring semantic consistency for cross-agent collaboration.

We adopt a block-wise random missingness strategy. Within each time window, modality corruption is independently applied to different blocks according to a predefined missing ratio. Missingness is simulated via feature masking or zeroing, mimicking real-world failures caused by occlusion, noise, or illumination changes. This temporally distributed corruption better reflects the dynamic nature of UAV sensing environments, where modality quality fluctuates over time.

To isolate and analyze the effect of collaborative compensation, we adopt a controlled setting: The target UAV experiences modality missingness or degradation; Neighboring UAVs are assumed to maintain reliable and complete modalities. This design allows us to evaluate the effectiveness of the compensation mechanism under a "single-node degradation, multi-node support" scenario, while avoiding confounding factors caused by simultaneous multi-node degradation.

Based on the above setup, Sections 5.2 and 5.3 evaluate the proposed method under Dual-UAV scenarios and Multi-UAV scenarios across varying modality missing rates and sample selection strategies.

5.2. Experiment A: Collaborative Compensation Under Dual-UAV Setting

We first evaluate the proposed method under a **dual-UAV setting**, consisting of two agents u_1 and u_2. This simplified topology allows us to isolate and analyze the effectiveness of the collaborative compensation mechanism.

To simulate modality missingness, we perform feature-level corruption by randomly zeroing out audio or visual features according to predefined missing ratios. This approximates real-world UAV scenarios where local observations may be partially unavailable or corrupted due to environmental factors.

Experiments are conducted under multiple buffer constraints, controlled via different sample retention (pruning) ratios. Based on the collaboratively compensated modality confidence computed for both u_1 and u_2, we further perform CC-MBS-based sample selection under varying selection ratios, aiming to validate the effectiveness of the proposed strategy.

 doi:10.20944/preprints202606.0287.v1

5.2.1. Baseline Performance Under Single-Modality Missingness

To establish a baseline, we first evaluate model performance under single-modality missingness in a single-UAV setting (i.e., without collaborative compensation).

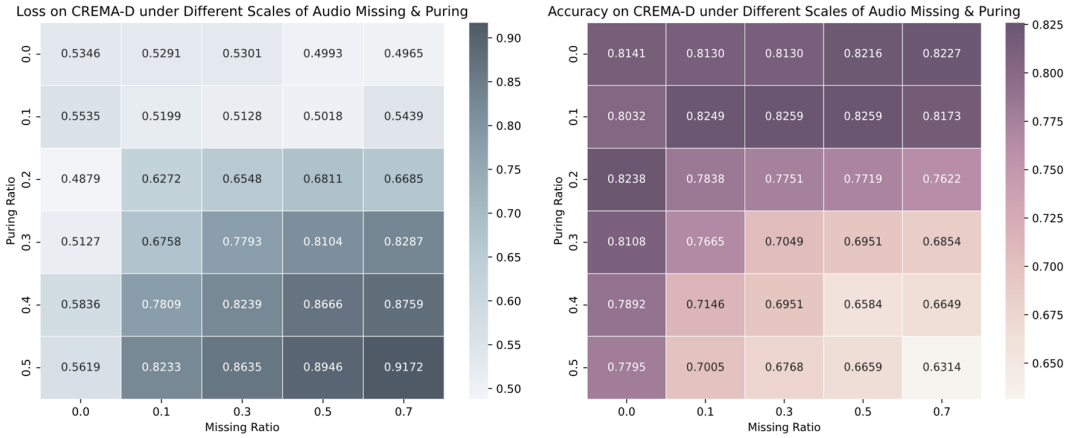
Specifically, we independently apply modality missingness to either audio or visual features with missing ratios of 10%, 30%, 50%, and 70%, and evaluate model performance under different sample selection ratios (10%, 20%, 30%, 40%, and 50%).

The training process consists of two stages:

- Pretraining stage (20 epochs): Used to compute the MBS score for each sample, simulating early-stage estimation of modality balance;
- Formal training stage (100 epochs): Conducted using the selected samples to allow the model to approach convergence.

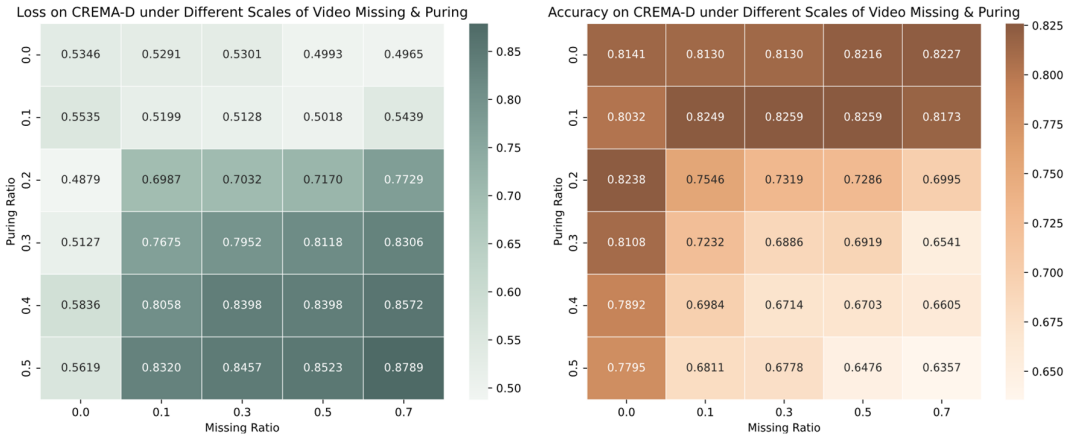
Final performance is evaluated on the test set in terms of classification accuracy.

This setup follows the protocol of prior work and reflects the practical scenario where sample valuation is performed at early stages of training. The results are presented in Figures 3–6.



(**a**) (**b**)

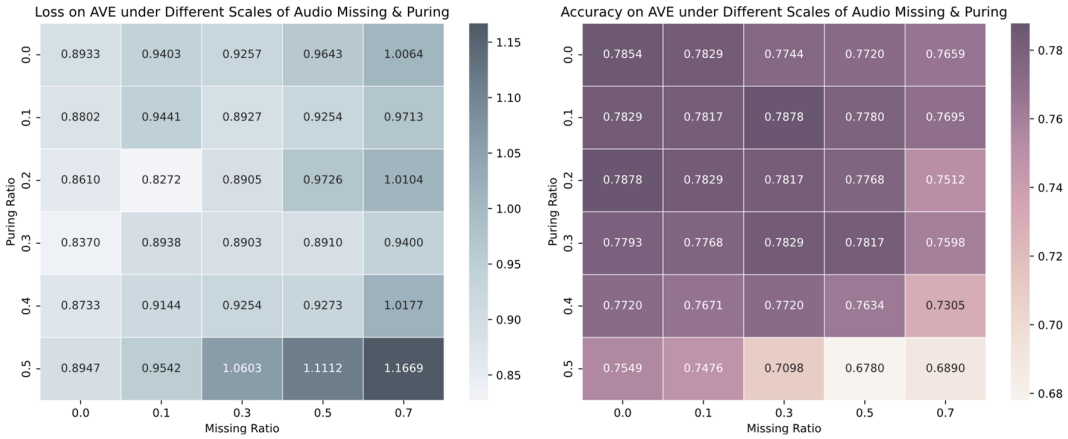
Figure 3. Performance on the CREMA-D dataset under different audio modality missing ratios and sample pruning ratios. (a) Training loss under audio modality missing ratios of 10%, 30%, 50%, and 70% with sample pruning ratios ranging from 10% to 50%; (b) Classification accuracy under the same settings.



(**a**) (**b**)

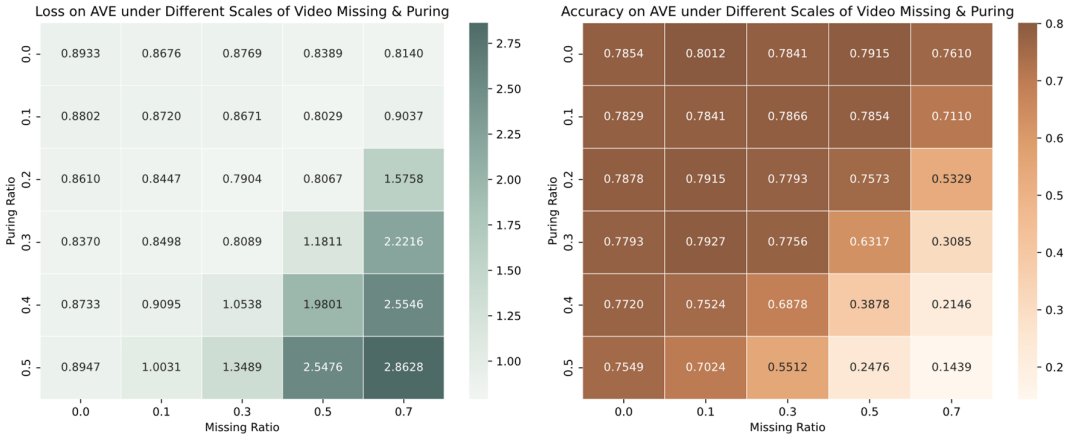
Figure 4. Performance on the CREMA-D dataset under different video modality missing ratios and sample pruning ratios. (a) Training loss under video modality missing ratios of 10%, 30%, 50%, and 70% with sample pruning ratios ranging from 10% to 50%; (b) Classification accuracy under the same settings.

 doi:10.20944/preprints202606.0287.v1



(**a**) (**b**)

Figure 5. Performance on the AVE dataset under different audio modality missing ratios and sample pruning ratios. (a) Training loss under audio modality missing ratios of 10%, 30%, 50%, and 70% with sample pruning ratios ranging from 10% to 50%; (b) Classification accuracy under the same settings.



(**a**) (**b**)

Figure 6. Performance on the AVE dataset under different video modality missing ratios and sample pruning ratios. (a) Training loss under video modality missing ratios of 10%, 30%, 50%, and 70% with sample pruning ratios ranging from 10% to 50%; (b) Classification accuracy under the same settings.

From the dataset perspective, the CREMA-D dataset, which focuses on emotion recognition, exhibits relatively balanced contributions from audio and visual modalities overall. However, a slight dominance of the audio modality can be observed. Specifically, under the same modality missing ratios, the overall performance trends for audio and visual missingness are similar, while in some low-missing-ratio settings, performance degradation caused by audio missingness is marginally more pronounced than that caused by visual missingness. This suggests that the audio modality plays a relatively more sensitive role in emotion representation.

In contrast, the AVE, which targets complex event recognition, demonstrates a clear visual dominance. Under audio missingness, the model performance degrades only slightly, whereas visual missingness leads to a more substantial performance drop.

Overall, as the modality missing ratio increases, the model performance exhibits a generally decreasing trend. Under certain mild missingness conditions, the model accuracy remains comparable to or slightly higher than the full-modality baseline. This phenomenon is consistent with a regularization-like effect at the modality level, where partial modality suppression may reduce over-reliance on dominant modalities and improve cross-modal generalization.

 doi:10.20944/preprints202606.0287.v1

These observations further motivate the introduction of collaborative compensation mechanisms under higher missingness conditions.

5.2.2. Evaluation of Collaborative Compensation Under Dual-UAV Setting

According to Eq. (20), the collaboratively compensated modality confidence depends on three key factors: the neighbor UAV's modality confidence $c_{j,t}^{m}$, the gating function $g_{i,t}^{m}$, and the compensation strength coefficient λ. For clarity, the compensation process can be expressed (under the dual-UAV setting) as:

$$\tilde{c}_{i,t}^{m} = c_{i,t}^{m} + \lambda \cdot g_{i,t}^{m} \cdot c_{j,t}^{m}, \tag{29}$$

where u_j denotes the neighboring UAV of u_i.

To evaluate the effectiveness of the proposed CC-MBS mechanism and the impact of different compensation strengths, we conduct ablation experiments under the dual-UAV topology.

We adopt a moderate sample pruning ratio of 20% as the default setting, and simulate modality missingness on both the CREMA-D and AVE datasets. Specifically:

- Modality missing ratios are set to 10%, 30%, 50%, and 70% for both audio and visual modalities;
- UAV u_1 is assigned modality missingness, where missing samples are distributed across the time window;
- UAV u_2 maintains complete modalities and serves as the collaborative reference;
- Both UAVs are assigned identical modality asynchrony factors to simulate realistic sensing and communication delays.

To analyze the effect of compensation strength, we vary the compensation strength coefficient $\lambda \in \{0.1, 0.2, \ldots, 1.0\}$, and evaluate the model performance on u_1 under the same training protocol (20 epochs pretraining + 100 epochs formal training). The results are reported in Figures 7 and 8.

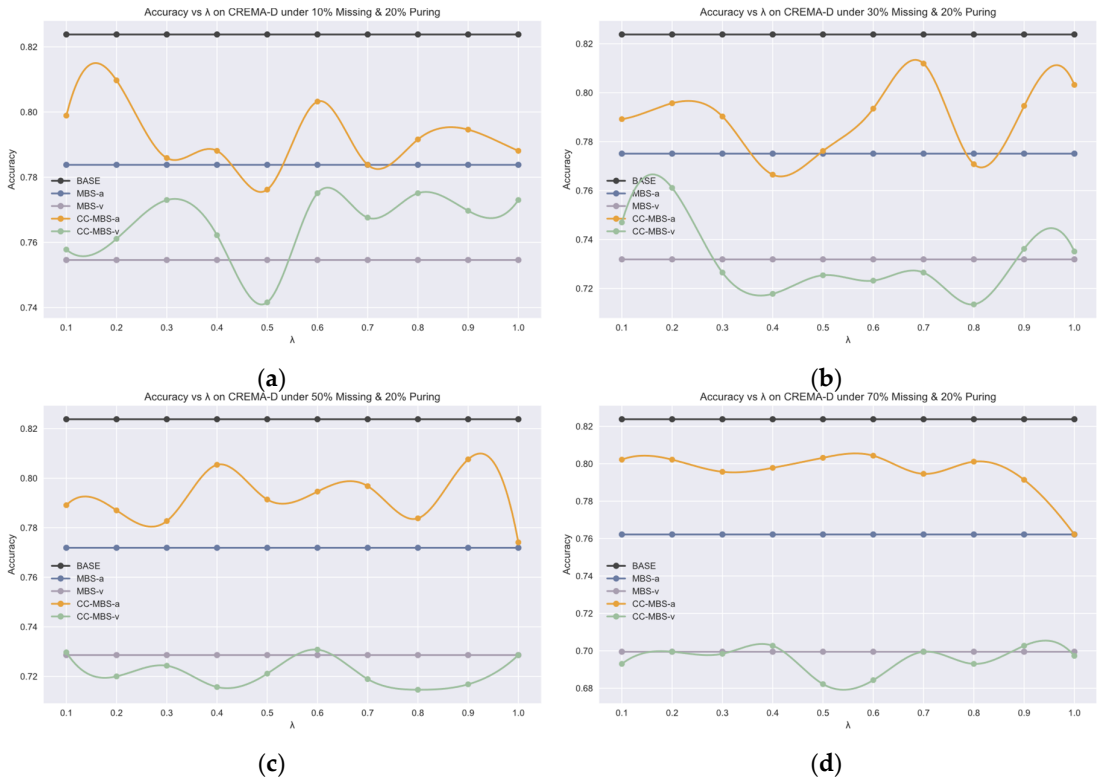


Figure 7. Performance of CC-MBS on the CREMA-D dataset under audio and visual modality missing conditions with different compensation strengths. (a) Accuracy under 10% modality missing ratio and 20% sample pruning; (b) Accuracy under 30% modality missing ratio and 20% sample pruning; (c) Accuracy under 50% modality missing ratio and 20% sample pruning; (d) Accuracy under 70% modality missing ratio and 20% sample pruning.

 doi:10.20944/preprints202606.0287.v1

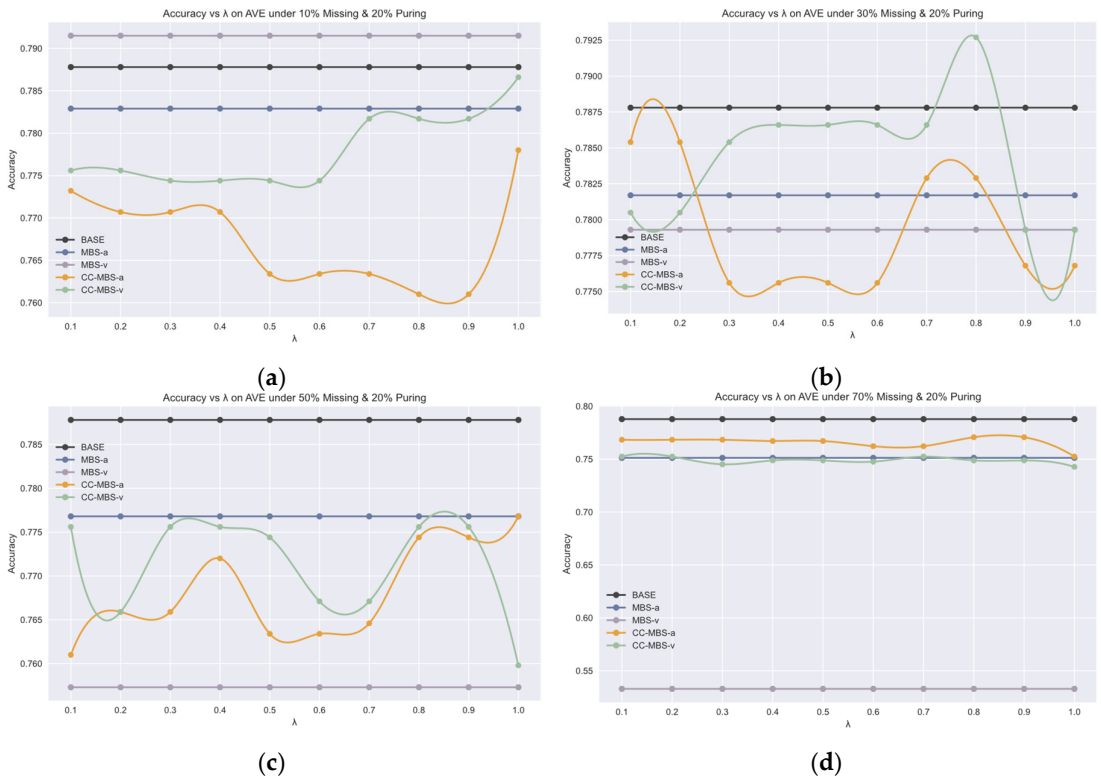


Figure 8. Performance of CC-MBS on the AVE dataset under audio and visual modality missing conditions with different compensation strengths. (a) Accuracy under 10% modality missing ratio and 20% sample pruning; (b) Accuracy under 30% modality missing ratio and 20% sample pruning; (c) Accuracy under 50% modality missing ratio and 20% sample pruning; (d) Accuracy under 70% modality missing ratio and 20% sample pruning.

Preliminary observations indicate that, under the collaborative compensation enabled by CC-MBS, the performance degradation caused by modality missingness is alleviated to a certain extent. Specifically, on the CREMA-D, the performance drop induced by audio missingness is reduced, while on the AVE, the degradation caused by visual missingness is similarly mitigated. These results provide empirical support for the effectiveness of the proposed CC-MBS mechanism, and are consistent with the modality dominance observations discussed in Section 5.2.1.

Further analysis shows that the compensation strength coefficient λ plays a critical role in balancing local modality confidence and collaborative neighborhood information. As the modality missing ratio increases, the value of λ corresponding to peak accuracy tends to shift toward higher values. This trend suggests that, when local modality quality deteriorates, the model increasingly relies on information from neighboring UAVs to maintain discriminative performance.

In addition, the effectiveness of compensation appears to vary across modalities depending on their relative importance in the task. When the dominant modality is missing, the local confidence decreases more significantly, triggering the compensation mechanism more frequently. In such cases, the contribution from neighboring UAVs becomes more prominent, leading to more noticeable performance improvements. In contrast, when non-dominant modalities are missing, the impact on overall performance is relatively smaller, and the benefit of compensation is correspondingly limited.

For example, in the CREMA-D dataset, where the audio modality plays a relatively important role in emotion recognition, compensation under audio missingness leads to relatively stable improvements, while visual missingness has a comparatively smaller impact. Conversely, in the AVE dataset, which exhibits stronger visual dominance, performance is more sensitive to visual missingness. Under audio missingness, the compensation mechanism is still activated; however, the contribution of neighboring audio information may be less consistent, which can result in less stable performance gains and more irregular trends with respect to λ.

 doi:10.20944/preprints202606.0287.v1

Overall, as the modality missing ratio increases, the effect of collaborative compensation becomes more pronounced. By adjusting the compensation strength coefficient λ, CC-MBS enables adaptive integration of local and collaborative information. Across different datasets and missingness conditions, the method consistently improves robustness, particularly under dominant-modality degradation, while yielding comparatively smaller gains for non-dominant modalities.

5.3. Experiment B: Collaborative Compensation Under Multi-UAV Setting

In this section, we extend our evaluation to **multi-UAV scenarios** to systematically assess the effectiveness and scalability of the proposed CC-MBS framework.

On one hand, we construct multi-UAV collaborative perception settings on both the CREMA-D and AVE datasets. For each dataset, modality missingness is applied to its dominant modality (i.e., audio for CREMA-D and visual for AVE), under varying missing ratios. This allows us to analyze the behavior of the collaborative compensation mechanism and the sample selection strategy under different swarm sizes and sensing conditions.

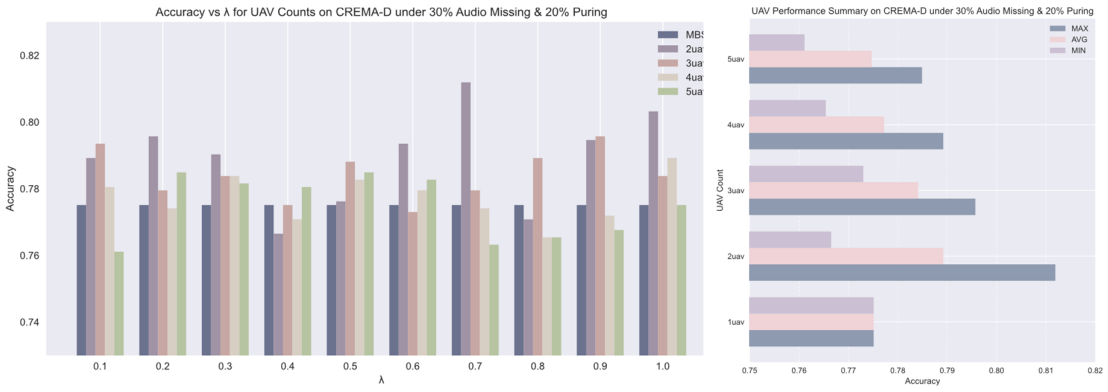
On the other hand, we reproduce representative methods in multimodal robust learning and sample selection under the same experimental settings, enabling a fair comparison to evaluate the effectiveness and advantages of CC-MBS in multi-UAV environments.

5.3.1. Baseline Performance of CC-MBS Under Multi-UAV Setting

Based on the observations from Sections 5.1 and 5.2, we conduct baseline evaluations under moderate modality missingness and light sample pruning settings. Specifically:

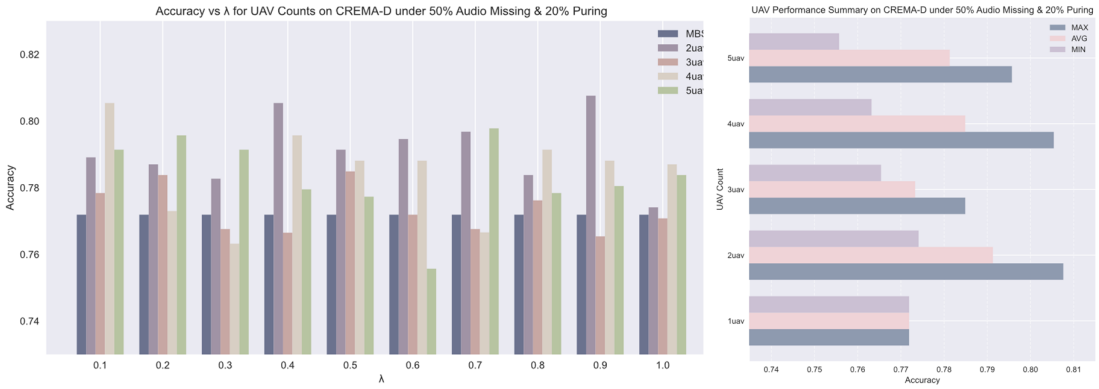
- Sample pruning ratio is fixed at 20%;
- Modality missing ratios are set to 30% and 50% on the dominant modality;
- Missingness is applied to UAV u_1, while other neighboring UAVs are assumed to have complete modalities.

To analyze scalability, we vary the number of UAVs from 1 to 5, and evaluate model performance under different compensation strengths $\lambda \in \{0.1, 0.2, \ldots, 1.0\}$. The final classification accuracy is measured under the same training protocol (20 epochs pretraining + 100 epochs formal training). The results are presented in Figures 9 and 10.



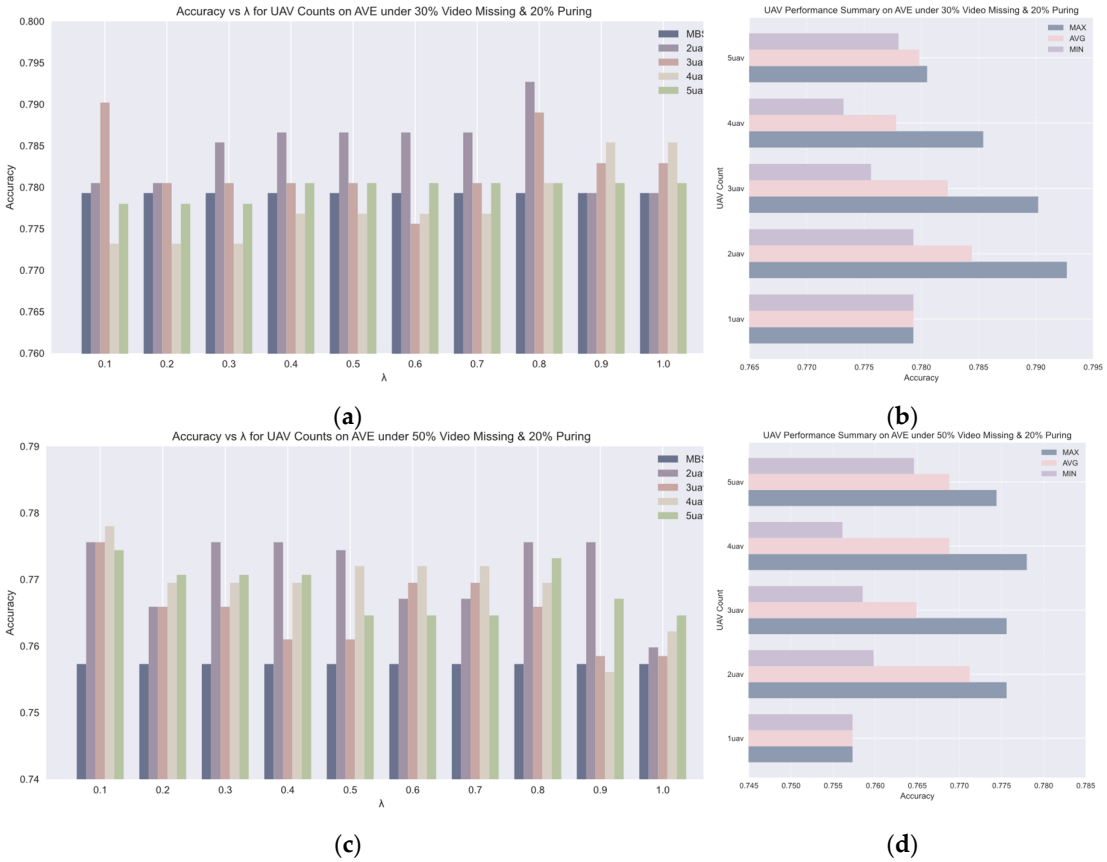
(**a**) (**b**)

 doi:10.20944/preprints202606.0287.v1



(**c**) (**d**)

Figure 9. Performance of CC-MBS in multi-UAV scenarios on the CREMA-D dataset under varying audio modality missing conditions. (a) Model accuracy with UAV counts from 1 to 5 under 30% audio modality missingness and 20% sample pruning, evaluated at different λ values; (b) Maximum, minimum, and average accuracy across UAV counts from 1 to 5 under the same conditions; (c) Model accuracy with UAV counts from 1 to 5 under 50% audio modality missingness and 20% sample pruning, evaluated at different λ values; (d) Maximum, minimum, and average accuracy across UAV counts from 1 to 5 under the same conditions.



(**a**) (**b**)

(**c**) (**d**)

Figure 10. Performance of CC-MBS in multi-UAV scenarios on the AVE dataset under varying video modality missing conditions. (a) Model accuracy with UAV counts from 1 to 5 under 30% video modality missingness and 20% sample pruning, evaluated at different λ values; (b) Maximum, minimum, and average accuracy across UAV counts from 1 to 5 under the same conditions; (c) Model accuracy with UAV counts from 1 to 5 under 50% video modality missingness and 20% sample pruning, evaluated at different λ values; (d) Maximum, minimum, and average accuracy across UAV counts from 1 to 5 under the same conditions.

 doi:10.20944/preprints202606.0287.v1

Preliminary observations show that under moderate to light modality missingness (30% and 50%), the CC-MBS mechanism significantly outperforms the MBS baseline in single-UAV settings, across multi-UAV swarms with 2–5 UAVs. This indicates that, under partial modality degradation but not complete failure, high-quality modality information from neighboring UAVs—transmitted as modality confidence (not full high-dimensional data)—can effectively compensate for the target UAV's perceptual insufficiency, thus improving the reliability of subsequent sample evaluation and selection.

However, upon further analysis of the results for different swarm sizes, considering the maximum, minimum, and average accuracy, we observe that the model performance does not continuously improve as the UAV node count increases. In some configurations, a slight decrease in performance is noted. This phenomenon can be attributed to two key factors:

- As the neighborhood expands, the modality information from different UAVs may differ in quality and distribution, introducing potential noise accumulation or redundant interference that weakens the compensation effectiveness;
- In the current collaborative mechanism, the contribution from each node is weighted primarily by modality confidence. When the neighborhood size grows large, weight distribution may become more diffuse, diluting the contribution of high-quality nodes and reducing the precision of compensation.

This trend is particularly evident under moderate modality missingness (30%), where smaller UAV swarms (2–3 UAVs) achieve a better balance between "information augmentation" and "noise control."

It is also worth noting that, at higher modality missing ratios (50%), the performance advantages seen in the 2–3 UAV swarm gradually diminish as the UAV node count increases. This does not contradict the previous inference, as the main reason lies in the fact that, under high missingness, the target node's modality representation becomes significantly incomplete, leading to instability in its role as a reference during collaborative compensation. As a result, cross-node fusion becomes more dependent on the "external injection" of neighborhood information. Despite the neighboring UAVs maintaining complete modalities, their observations remain susceptible to viewpoint differences, environmental disturbances, and cross-node distribution shifts. As the number of collaborative nodes increases, this added variability may introduce inconsistency, thereby weakening the effectiveness of the compensation.

Overall, these experimental results confirm that the CC-MBS mechanism can effectively improve model performance in multi-UAV swarms (within moderate sizes). However, the performance gains are not linearly correlated with the number of UAVs, underscoring the importance of reasonably controlling the collaborative range to achieve stable performance improvements.

5.3.2. Performance Comparison of Modality Missingness Compensation

To further validate the performance of CC-MBS in modality missingness compensation within a UAV swarm, we compare it against the representative ShaSpec method [16], which employs shared-specific feature modeling. ShaSpec explicitly models shared and modality-specific representations, decomposing multimodal features into "cross-modal consistent information" and "modality-specific information." During training, the shared representation enhances cross-modal consistency, while the specific representation preserves independent features for each modality. As a result, ShaSpec compensates for missing modalities by relying on the shared representation, enabling the model to maintain some discriminative ability even when a modality is missing. This approach can be viewed as a representation-layer compensation strategy, where the model becomes more tolerant to missing inputs without relying on external sources of information.

In contrast, the CC-MBS method proposed in this paper focuses on data selection and collaborative compensation. Instead of relying on complete high-dimensional data or sample features, CC-MBS recalculates sample value by sharing modality confidence across UAV nodes. This enables dynamic sample selection, improving the quality of training samples at the data level. The

 doi:10.20944/preprints202606.0287.v1

two methods differ significantly in their design philosophy and operational mechanisms, as summarized in Table 1.

Table 1. Comparison between CC-MBS and ShaSpec methods.

Comparison Dimension	ShaSpec	CC-MBS
Core Idea	Shared-specific feature decomposition	Modality confidence-driven collaborative compensation
Information Source	Single node internal	Across UAV nodes
Handling Modality Missingness	Relies on shared features to fill gaps	Compensation via neighboring UAVs
Compensation Layer	Feature representation layer	Data selection layer + cross-node information layer
Modality Quality Consideration	No (implicit modeling)	Yes (explicit modeling of modality confidence)
Sensitivity to Noise	Ordinary	Reduced (sample selection)

Since ShaSpec is originally designed for single-node multimodal learning, we extend it with three common multi-node model parameter aggregation strategies:

- AVG aggregation: Average the model parameters from multiple nodes to obtain the global model for performance evaluation;
- PFM aggregation: Weight the aggregation based on each node's local performance and compute a weighted average for the global model;
- POW aggregation: Assign higher aggregation weights to nodes with stronger local performance (e.g., 60%/40% split for 2 UAVs) when calculating the global model.

We conduct performance tests using light sample pruning (20%) across UAV swarm sizes from 1 to 5 UAVs. The results for the CREMA-D and AVE datasets, under different dominant modality missingness ratios (30% and 50%), are presented in Figures 11 and 12.

On the CREMA-D and AVE datasets, when no compensation is applied (i.e., single UAV), CC-MBS degrades to the MBS baseline method, showing significantly lower performance than the ShaSpec method due to the lack of modality missingness compensation. However, in multi-UAV settings, CC-MBS consistently outperforms all three parameter aggregation strategies (AVG, PFM, and POW) derived from the extended ShaSpec method. This result indicates that data-level methods like CC-MBS, which rely on collaborative compensation and dynamic sample selection, exhibit superior robustness in the presence of modality missingness compared to traditional model-parameter fusion strategies.

Furthermore, within the extended ShaSpec aggregation methods, PFM and POW generally outperform AVG, suggesting that weighting node performance helps mitigate inconsistencies in multi-node model fusion. However, these methods remain limited by their reliance on model-parameter post-fusion and lack explicit modeling of modality reliability, leading to relatively smaller performance gains. In contrast, CC-MBS addresses this issue by explicitly modeling modality quality through collaborative compensation, significantly improving robustness in multimodal missingness scenarios.

 doi:10.20944/preprints202606.0287.v1

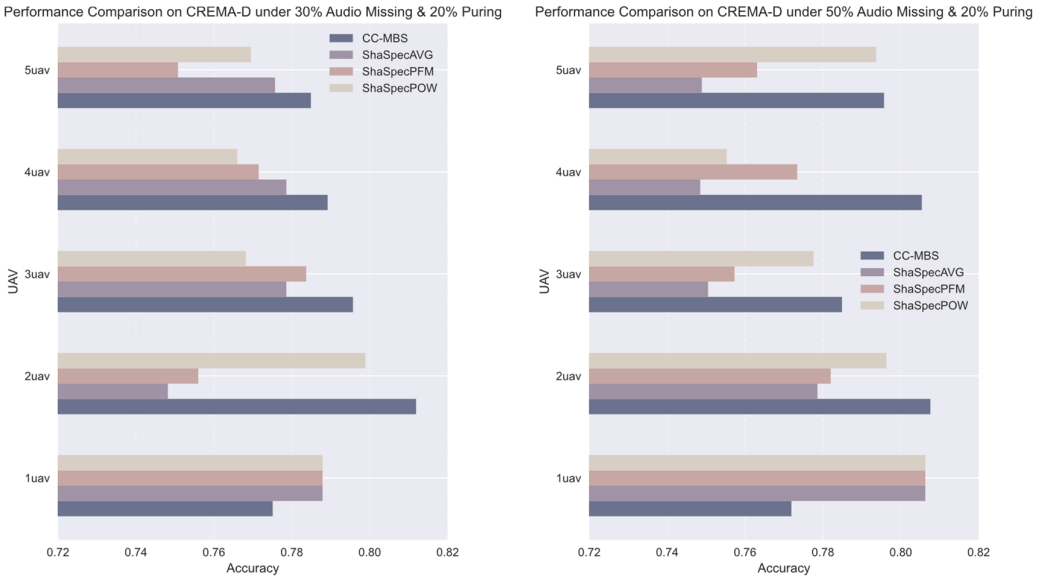


Figure 11. Performance comparison on CREMA-D between CC-MBS and ShaSpec under different audio modality missing ratios and 20% sample pruning. (a) Accuracy of CC-MBS and ShaSpec with three aggregation strategies (AVG, PFM, POW) across UAV swarm sizes from 1 to 5 under 30% audio modality missingness; (b) Accuracy of CC-MBS and ShaSpec with three aggregation strategies (AVG, PFM, POW) across UAV swarm sizes from 1 to 5 under 50% audio modality missingness.

Figure 12. Performance comparison on AVE between CC-MBS and ShaSpec under different video modality missing ratios and 20% sample pruning. (a) Accuracy of CC-MBS and ShaSpec with three aggregation strategies (AVG, PFM, POW) across UAV swarm sizes from 1 to 5 under 30% video modality missingness; (b) Accuracy of CC-MBS and ShaSpec with three aggregation strategies (AVG, PFM, POW) across UAV swarm sizes from 1 to 5 under 50% video modality missingness.

In summary, this set of experiments demonstrates that, in multi-UAV environments, traditional model-parameter aggregation methods fail to fully exploit cross-node modality compensation. CC-MBS, by combining modality quality modeling with collaborative compensation, provides more robust and stable performance across different datasets and missingness conditions. Additionally,

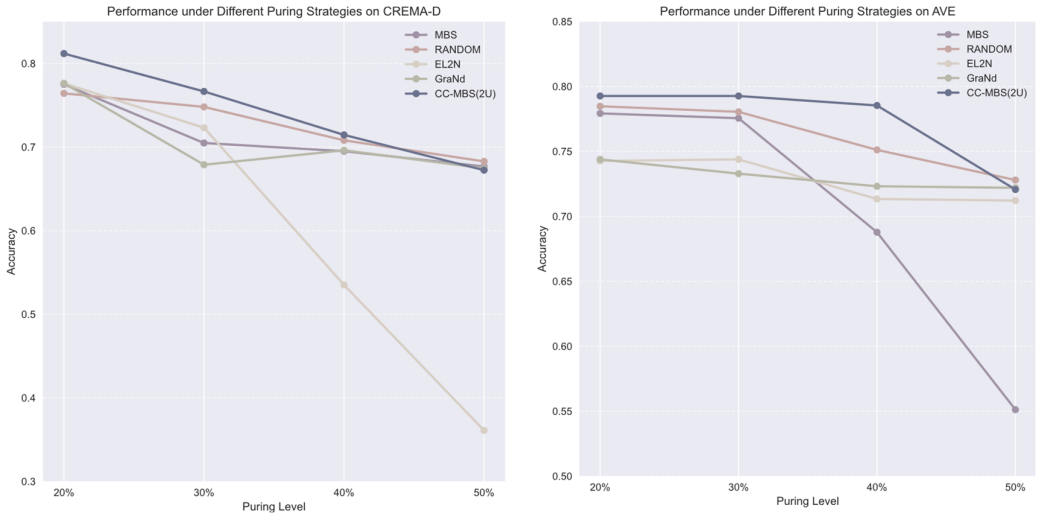
 doi:10.20944/preprints202606.0287.v1

CC-MBS achieves this with lightweight modality confidence sharing, making it a more efficient and scalable solution in distributed multi-node environments.

5.4. Experiment C: Evaluation of Sample Selection

Based on the results from Sections 5.2 and 5.3, it is evident that CC-MBS achieves optimal collaborative compensation in small-scale UAV swarms consisting of 2–3 nodes. In this section, we further evaluate the sample selection capability of CC-MBS in a 2-UAV setup under mild dominant modality missingness (30%), and compare it against representative data selection methods: EL2N, GraNd [22], and random sampling (RANDOM). The experimental results are shown in Figure 13.

Across both the AVE and CREMA-D datasets, CC-MBS consistently demonstrates superior or competitive performance across different sample pruning ratios (20%–50%), with its advantage becoming more pronounced at higher pruning ratios. This indicates that, under mild dominant modality missingness, CC-MBS provides enhanced robustness, maintaining stable model performance even when a substantial portion of the training data is pruned.



(**a**) (**b**)

Figure 13. Performance of different pruning strategies under modal missingness in a 2-UAV swarm with 30% audio modality missingness. (a) Accuracy on the CREMA-D dataset across varying sample pruning ratios (20%–50%) for CC-MBS, MBS, EL2N, GraNd, and RANDOM strategies. (b) Accuracy on the AVE dataset under the same experimental settings.

The MBS method exhibits significant limitations under modality missing conditions, particularly at high pruning ratios. Its sample scoring mechanism relies heavily on single-node modality information; when the dominant modality is partially missing or degraded—especially in datasets with poor modality balance like AVE—this scoring process is directly affected. Consequently, the evaluation of sample importance becomes biased. Under high pruning ratios, this bias is amplified, potentially leading to the removal of genuinely high-value samples and degrading the model's learning capability.

Similarly, EL2N and GraNd also show limitations under modality missing conditions. Both methods rely on dynamic training signals (e.g., loss values or gradient norms) to assess sample importance. However, when modalities are partially missing or degraded, the model's responses to input samples are disturbed, rendering loss and gradient signals less reliable indicators of true sample value. In particular, samples with high semantic content may generate abnormally high losses or unstable gradients under dominant modality missingness, causing them to be misclassified as "difficult but low-value" samples.

 doi:10.20944/preprints202606.0287.v1

Interestingly, RANDOM sampling exhibits relatively stable performance at high pruning ratios and, in some cases, outperforms EL2N and GraNd. This suggests that sample selection strategies relying solely on a single evaluation signal can introduce additional bias in multimodal missingness scenarios. Nevertheless, while RANDOM maintains stability, its lack of effective modeling of sample value results in overall performance that remains substantially lower than CC-MBS.

In contrast, CC-MBS leverages a cross-node collaborative compensation mechanism, allowing the target node to correct modality confidence using neighborhood information even when modalities are partially degraded. This provides a more reliable basis for sample evaluation during the selection stage. This single-node to multi-node extension effectively mitigates the information insufficiency problem encountered by MBS under incomplete modalities, enabling the model to maintain strong performance under the dual constraints of high pruning ratios and modality missingness.

Overall, this set of experiments demonstrates that, in small-scale UAV collaborative scenarios, CC-MBS not only enhances modality robustness through collaborative compensation but also effectively identifies high-value samples during data selection. This enables superior performance under limited data budgets, further validating its potential for deployment in resource-constrained distributed environments.

5.5. Comprehensive Analysis

The experimental results discussed above indicate that, across varying modality missing ratios and diverse task scenarios, CC-MBS consistently demonstrates stable performance improvements and mitigates the performance degradation caused by incomplete modalities. In particular, in small-scale collaborative scenarios consisting of 2–3 UAV nodes, the method effectively balances cross-node information supplementation and noise introduction, resulting in significant performance gains.

Under identical collaboration conditions, CC-MBS only requires the exchange of low-dimensional modality confidence vectors between nodes. Its communication overhead scales linearly with the number of modalities, making it more lightweight compared to feature-level or model parameter-level information exchange. Furthermore, comparisons with other sample selection methods indicate that, even under high sample pruning ratios and modality missing conditions, CC-MBS maintains relatively stable data utilization efficiency and model performance.

A deeper analysis of the experimental phenomena reveals that the effectiveness of the collaborative compensation mechanism is not only related to the neighborhood size but also influenced by cross-node information consistency and weight allocation strategies. Under moderate or low missing ratios, neighborhood information primarily serves as a supplementary source, improving the reliability of sample evaluation. At higher missing ratios, however, as the number of nodes increases, distributional discrepancies and accumulated noise across nodes may interfere with compensation, causing performance gains to plateau or fluctuate.

It is important to note that the experiments in this chapter were conducted under a controlled setting in which single-node modality degradation occurs while neighboring nodes remain relatively reliable. This setup is representative of scenarios where local nodes are affected by occlusion or transient disturbances, while neighboring nodes maintain stable sensing capabilities. In more general cases, multiple nodes may simultaneously experience varying degrees of modality degradation, in which case the effectiveness of collaborative compensation will depend further on node-quality distribution and information consistency. Modeling and optimizing for such complex scenarios will be a focus of future work.

6. Conclusions

This work addresses the challenges of modality missingness, signal degradation, and cross-modal asynchrony in UAV swarms operating in complex environments. We propose a Compensation-based Collaborative Modality-Balanced Sample Selection (CC-MBS) framework, which models modality confidence to capture the reliability of multimodal data and leverages a

 doi:10.20944/preprints202606.0287.v1

neighborhood collaborative compensation mechanism to correct incomplete local modalities using cross-node information. This enables more robust sample value evaluation and selection during incremental learning.

Experimental results demonstrate that CC-MBS consistently achieves stable and robust performance across varying modality missing ratios and multi-UAV collaborative scenarios. Particularly in small-scale collaborative settings, it effectively enhances data utilization efficiency and mitigates performance degradation. Further analysis indicates that the effectiveness of collaborative compensation is jointly influenced by neighborhood size and information consistency. Future work will focus on developing adaptive collaborative strategies for complex multi-node degradation scenarios, aiming to extend the applicability of CC-MBS to larger-scale UAV swarms.

Author Contributions: Conceptualization, Y.X. and B.C.; methodology, Y.X.; software, Y.X.; validation, Y.X. and Y.C.; formal analysis, Y.X.; investigation, Y.X.; resources, Y.C. and Z.X.; data curation, Y.X.; writing—original draft preparation, Y.X.; writing—review and editing, B.C., F.H. and Y.C.; visualization, Y.X.; supervision, B.C.; project administration, B.C.; funding acquisition, B.C. and Z.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China under Grant No. 62402221, the Natural Science Foundation of Jiangsu Province under Grant No. BK20241379, the China Postdoctoral Science Foundation under Grant No. 2025M774281, Jiangsu Funding Program for Excellent Postdoctoral Talent.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author(s).

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

UAV	Unmanned aerial vehicle
MBS	Modality Balance Score
CC-MBS	Compensatory Collaboration Modality-Balanced (Sample) Selection framework
AVG	Average Aggregation Method
PFM	Performance-based Aggregation Method
POW	Power-based Aggregation Method

References

1. Xu Y, Chen B, Hu F, et al. MBS: A Modality-Balanced Strategy for Multimodal Sample Selection[J]. Machine Learning and Knowledge Extraction, 2026, 8(1): 17.
2. Zhang H, Hanzo L. Federated learning assisted multi-UAV networks[J]. IEEE Transactions on Vehicular Technology, 2020, 69(11): 14104-14109.
3. He G, Li C, Song M, et al. A hierarchical federated learning incentive mechanism in UAV-assisted edge computing environment[J]. Ad Hoc Networks, 2023, 149: 103249.
4. Tong Z, Wang J, Hou X, et al. Blockchain-based trustworthy and efficient hierarchical federated learning for UAV-enabled IoT networks[J]. IEEE Internet of Things Journal, 2024, 11(21): 34270-34282.
5. Wang Z, Cheng P, Chen M, et al. Drones help drones: A collaborative framework for multi-drone object trajectory prediction and beyond[J]. Advances in Neural Information Processing Systems, 2024, 37: 64604-64628.
6. Lin Z, Chen W, Jin X, et al. MCOP: Multi-UAV Collaborative Occupancy Prediction[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2025: 27242-27251.

 doi:10.20944/preprints202606.0287.v1

7. Bocheng Z, Mingying H U O, Zheng L I, et al. Graph-based multi-agent reinforcement learning for collaborative search and tracking of multiple UAVs[J]. Chinese Journal of Aeronautics, 2025, 38(3): 103214.
8. Havaei M, Guizard N, Chapados N, et al. Hemis: Hetero-modal image segmentation[C]//International conference on medical image computing and computer-assisted intervention. Cham: Springer International Publishing, 2016: 469-477.
9. Ma M, Ren J, Zhao L, et al. Smil: Multimodal learning with severely missing modality[C]//Proceedings of the AAAI conference on artificial intelligence. 2021, 35(3): 2302-2310.
10. Zhao J, Li R, Jin Q. Missing modality imagination network for emotion recognition with uncertain missing modalities[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 2608-2618.
11. Poklukar P, Vasco M, Yin H, et al. Geometric multimodal contrastive representation learning[C]//International Conference on Machine Learning. PMLR, 2022: 17782-17800.
12. Lee K, Lee S, Hahn S, et al. Learning missing modal electronic health records with unified multi-modal data embedding and modality-aware attention[C]//Machine Learning for Healthcare Conference. PMLR, 2023: 423-442.
13. Liu H, Wei D, Lu D, et al. M3AE: multimodal representation learning for brain tumor segmentation with missing modalities[C]//Proceedings of the AAAI conference on artificial intelligence. 2023, 37(2): 1657-1665.
14. Lin R, Hu H. Missmodal: Increasing robustness to missing modality in multimodal sentiment analysis[J]. Transactions of the Association for Computational Linguistics, 2023, 11: 1686-1702.
15. Li M, Yang D, Liu Y, et al. Toward robust incomplete multimodal sentiment analysis via hierarchical representation learning[J]. Advances in Neural Information Processing Systems, 2024, 37: 28515-28536.
16. Wang H, Chen Y, Ma C, et al. Multi-modal learning with missing modality via shared-specific feature modelling[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 15878-15887.
17. Lee Y L, Tsai Y H, Chiu W C, et al. Multimodal prompting with missing modalities for visual recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 14943-14952.
18. Hu L, Shi T, Feng W, et al. Deep correlated prompting for visual recognition with missing modalities[J]. Advances in Neural Information Processing Systems, 2024, 37: 67446-67466.
19. Li M, Yang D, Lei Y, et al. A unified self-distillation framework for multimodal sentiment analysis with uncertain missing modalities[C]//Proceedings of the AAAI conference on artificial intelligence. 2024, 38(9): 10074-10082.
20. Cao H, Cooper D G, Keutmann M K, et al. CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset[J]. IEEE Transactions on Affective Computing, 2014, 5(4): 377-390.
21. Tian Y, Shi J, Li B, et al. Audio-Visual Event Localization in Unconstrained Videos[C]//Proceedings of the European Conference on Computer Vision (ECCV). Cham: Springer, 2018: 247-263.
22. Paul M, Ganguli S, Dziugaite G K. Deep learning on a data diet: Finding important examples early in training[C]//Advances in Neural Information Processing Systems (NeurIPS). 2021, 34: 20596-20607.