Frugal Federated Learning for Violence Detection: A Comparison of LoRA-Tuned VLMs and Personalized CNNs

Sébastien Thuau*, Siba Haidar*, Ayush Bajracharya[†], Rachid Chelouah[†]
*esieaLab, ESIEA, Paris, France

Email: thuau@et.esiea.fr, siba.haidar@esiea.fr

†ETIS Laboratory, CNRS, UMR8051, University of CY Cergy, Paris, France
Email: ayush.bajracharya@etu.cyu.fr, rc@cy-tech.fr

Abstract—We examine frugal federated learning approaches to violence detection by comparing two complementary strategies: (i) zero-shot and federated fine-tuning of vision-language models (VLMs), and (ii) personalized training of a compact 3D convolutional neural network (CNN3D). Using LLaVA-7B and a 65.8M parameter CNN3D as representative cases, we evaluate accuracy, calibration, and energy usage under realistic non-IID settings.

Both approaches exceed 90% accuracy. CNN3D slightly outperforms Low-Rank Adaptation(LoRA)-tuned VLMs in ROC AUC and log loss, while using less energy. VLMs remain favorable for contextual reasoning and multimodal inference. We quantify energy and CO_2 emissions across training and inference, and analyze sustainability trade-offs for deployment.

To our knowledge, this is the first comparative study of LoRA-tuned vision-language models and personalized CNNs for federated violence detection, with an emphasis on energy efficiency and environmental metrics.

These findings support a hybrid model: lightweight CNNs for routine classification, with selective VLM activation for complex or descriptive scenarios. The resulting framework offers a reproducible baseline for responsible, resource-aware AI in video surveillance, with extensions toward real-time, multimodal, and lifecycle-aware systems.

Index Terms—Federated Learning, Non-IID Data, Vision-Language Models, LoRA Fine-Tuning, Violence Detection, Video Surveillance, Energy Efficiency, Multimodal Artificial Intelligence

I. INTRODUCTION

Video surveillance systems increasingly rely on deep learning models for detecting and analyzing violent scenes in public spaces. However, centralized pipelines raise critical privacy concerns, especially when sensitive video data is collected and processed off-device. In parallel, the computational and environmental costs of deploying large models at scale have drawn scrutiny from both researchers and regulators.

This paper is part of the **DIVA** initiative, which stands for *Decentralized Intelligence for Visual Awareness*. DIVA investigates sustainable, privacy-preserving AI methods for realworld video surveillance, with a focus on federated learning, multimodal reasoning, and energy-aware deployment.

Federated learning (FL) offers a promising solution by allowing models to be trained across multiple edge de-

vices without sharing raw data, using secure aggregation and privacy-preserving protocols. Yet, FL alone does not resolve the challenge of model size and efficiency—particularly for multimodal architectures like vision-language models (VLMs), which may carry billions of parameters. We adopt "Vision-Language Models (VLMs)" as a general term for architectures that jointly process visual and textual inputs. This usage reflects the most common naming convention in the literature [1]–[3].

In this study, we address two open questions: (1) whether large-scale VLMs can be frugally and effectively fine-tuned in federated non-IID settings, and (2) whether their deployment costs—especially in terms of energy and emissions—are justified when compared to lightweight CNNs that already perform competitively in violence classification tasks.

To this end, we conduct a systematic evaluation of centralized and federated learning strategies for video-based violence detection. We contrast zero-shot and Low-Rank Adaptation(LoRA)-adapted VLMs (e.g., LLaVA) with a compact 3D CNN trained under personalized federated learning. Our experiments quantify not only classification performance but also energy consumption and CO₂ emissions—providing practical insights for sustainable, privacy-aware deployment.

II. RELATED WORK

A. Privacy-Preserving Architectures for Video Surveillance

Federated learning has been widely recognized as a key strategy to enhance privacy in video surveillance, where raw video data can contain highly sensitive personal information. Unlike centralized machine learning pipelines that require uploading raw footage to external servers, federated systems retain data locally and transmit only model updates or abstracted features. Several recent studies affirm this advantage.

Jiang et al. [4] and Singh et al. [5] describe architectures in which user or edge devices train local models, never exposing raw data, while secure aggregation protocols ensure privacy during synchronization. These are supported by large-scale surveys such as Abdulrahman et al. [6] and Bellavista et al. [7], who discuss differential privacy, homomorphic encryption,

ephemeral updates, and multi-party computation as layered privacy mechanisms.

Vyas and Torrens [8] propose discarding raw video content post-processing and transmitting only AI-extracted motifs, which aligns with our work's goal of frugal, privacy-aware deployment.

Overall, this body of work establishes federated learning as a privacy-enhancing alternative to centralized surveillance, with improved scalability, reduced network load, and multilayered security—yet none explicitly address the cost-benefit trade-offs of federated fine-tuning for VLMs, which is the core contribution of this paper.

B. LoRA Fine-Tuning in Federated Learning

LoRA-based fine-tuning is theoretically well-suited to federated learning due to its parameter-efficient design, enabling significant reductions in communication and computation costs—critical factors in decentralized, resource-constrained environments.

C. Energy-Efficient Multimodal AI Models

Recent advances in personalized federated learning have introduced adapter-based fine-tuning as an efficient alternative to full model updates, particularly in low-resource or privacy-sensitive scenarios. LoRA (Low-Rank Adaptation) has emerged as a key mechanism in this context, enabling reduced communication and memory usage while maintaining competitive accuracy.

Below, we outline the key characteristics and empirical outcomes of recent LoRA-based federated learning frameworks:

- FLoCoRA [9]: Designed for vision models, it combines aggregation-agnostic LoRA adapters with affine quantization. This method achieves up to 63.9x reduction in communication with less than 4% accuracy loss.
- **LoRA-FAIR** [10]: Integrates server-side correction and refined initialization strategies. It delivers performance improvements without incurring additional communication costs, outperforming prior state-of-the-art methods.
- FLORA [3]: Applies rank-2 LoRA adapters to CLIP in federated settings. Reported gains include 30% improvement in accuracy, 34.7× training speedup, and 2.47× reduction in memory usage across 16 vision datasets. FLORA represents the closest multimodal approach to our use case of federated surveillance, but it has not been evaluated on real-world video violence detection tasks.
- FedLoRA [11]: Addresses client heterogeneity using a shared set of homogeneous LoRA adapters. Demonstrates 11.8x computational efficiency and a +1.35% accuracy gain.
- FDLoRA [12]: Employs a dual-adapter architecture
 that fuses global and personalized LoRA components.
 Achieves the highest mean accuracy across evaluated settings, emphasizing local adaptability without sacrificing
 global coherence.
- HAFL [13]: Explores adaptive-rank tuning of LoRA adapters, which provides low communication overhead

while avoiding performance degradation, making it suitable for highly constrained environments.

These contributions can be classified into two methodological families, Fig. 1:

- Aggregation-enhancing strategies: FLORA and LoRA-FAIR improve server-side coordination, particularly during model aggregation.
- Adapter-level innovations: FDLoRA, FLoCoRA, HAFL, and FedLoRA emphasize efficient local tuning through quantization, personalization, and modular design.

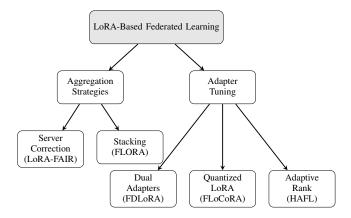


Fig. 1. Taxonomy of LoRA-based federated learning approaches. Methods are grouped into (left) aggregation-side coordination (*LoRA-FAIR*, *FLORA*) and (right) client-side adapter tuning (*FDLoRA*, *FLoCoRA*, *HAFL*). Our implementation follows the adapter-tuning path with standard LoRA modules on VLMs. The diagram is conceptual (no ranking implied).

Despite promising results, none of the reviewed studies evaluate LoRA-based federated fine-tuning of VLMs for violence detection in realistic video surveillance settings, nor do they compare their cost-effectiveness against lightweight CNNs under identical non-IID scenarios. While several frameworks demonstrate strong potential for efficiency and personalization, their application to federated multimodal fine-tuning in video surveillance remains largely untested. In contrast to prior work focused on general vision or multimodal tasks, our study fills this gap by applying federated LoRA fine-tuning to VLMs in video-based violence detection—benchmarking both performance and environmental cost against a lightweight CNN baseline under non-IID conditions.

Beyond algorithmic considerations, the sustainability of AI models—particularly their energy and emissions footprint—is increasingly central to deployment decisions.

D. Environmental Impact Assessment

The environmental impact of AI has become a major societal concern, particularly as deep learning models are increasingly deployed in large-scale systems such as violence detection. A critical issue is the lack of transparency regarding the energy consumption and carbon emissions of various model architectures. This gap in knowledge hinders efforts to optimize AI systems for sustainability. Initiatives like the *EU AI Act* and *AI Summit* emphasize the need for robust regulatory

frameworks to address these challenges, ensuring that AI technologies are developed and deployed with consideration for their environmental footprint. In France, projects like *The Shift Project*, supported by the *ADEME* (French Environment and Energy Management Agency), and *AFNOR SPEC* standards (e.g. *AFNOR SPEC 2314*), are actively working on defining best practices and guidelines for reducing the carbon impact of AI, promoting energy-efficient AI model development and fostering sustainable innovation in the field.

This paper responds to these emerging regulatory and scientific demands by quantifying the energy use and emissions of centralized and federated fine-tuning strategies for VLMs and benchmarking them against frugal CNN baselines, thus informing future deployment choices in sustainable surveillance.

III. METHODOLOGY

A. Overview

Our study evaluates the trade-offs between personalized federated learning of lightweight CNN models [16], [17] and LoRA-based fine-tuning of vision-language models (VLMs) in decentralized video surveillance. We perform three experiments on a merged dataset (RWF-2000 + RLVS) under various training and inference settings, measuring performance and energy efficiency. We simulate realistic federated non-IID scenarios with controlled heterogeneity in video content and label distribution.

B. Datasets and Client Partitioning

We construct a combined dataset of 4000 labeled videos by merging the RWF-2000 [18] and RLVS [19] datasets, each contributing 2000 videos with binary labels (violent vs. non-violent). An 80/20 stratified split is used for training and validation.

To simulate federated non-IID conditions and contextual heterogeneity, we partition the data across 10 clients: the first five clients receive only RWF-2000 videos, and the remaining five only RLVS. This structure reflects real-world variation in surveillance sources (e.g., shopping mall vs. subway) and enforces label and domain skew. For one experiment, we additionally use a Dirichlet distribution with concentration parameter $\alpha=1$ to vary the degree of label imbalance within clients.

C. Federated Learning Setup

All federated training uses the FedAvg algorithm [20], with 20 communication rounds. Experiments are conducted using a server-side simulation of federated learning: instead of deploying across physical edge devices, we emulate 10 independent clients on a single GPU server. Each client maintains isolated data partitions and local update procedures, and communication dynamics are simulated to reflect realistic federated behavior.

We simulate two levels of partial participation:

• 50% client fraction: 5 out of 10 clients randomly selected per round.

• 30% client fraction: 3 out of 10 clients randomly selected per round.

Partial participation is used to mirror bandwidth and availability constraints common in real-world FL deployments, while reducing cumulative training cost.

D. LoRA-Based Fine-Tuning: Centralized vs. Federated

We apply LoRA [21] fine-tuning to the <code>llava-hf/LLaVA-NeXT-Video-7B-hf</code>, a multimodal vision-language model using CLIP as the visual encoder. Videos are converted into representative frames (default: 24 frames per clip), which are passed through CLIP's visual backbone. Prompts such as "Analyze the video. Is this a fight scene? Answer yes or no." are used for both training and zero-shot evaluation.

Fine-tuning is conducted under two settings:

- Centralized LoRA: LLaVA is fine-tuned on the full training set using a standard supervised objective.
- Federated LoRA: The same model is fine-tuned using federated LoRA with 4-bit quantization. Each client updates local LoRA adapters while keeping base model weights frozen. Communication involves only adapter weights.

LoRA adapters are initialized with rank 8 and scaling factor $\alpha=32$ using the Hugging Face peft library. The adapter layers are inserted into attention blocks of the LLaVA architecture.

E. Personalized CNN3D with Parameter Decoupling

For the baseline, we implement a custom 3D CNN model with approximately 66 million parameters. Instead of standard FedAvg, we apply personalized federated learning (PFL) using parameter decoupling. Specifically, all layers after spatiotemporal feature extraction are decoupled and trained locally per client, while shared layers are aggregated globally. This balances generalization with client-specific specialization, which is particularly valuable under domain heterogeneity (RWF vs. RLVS).

F. Evaluation Protocol: Performance and Environmental Metrics

We assess each method across three complementary experiments:

- 1) **Exp. 1: Zero-shot inference with VLMs.** Pretrained models such as LLaVA, Qwen-VL, and Flamingo are evaluated directly without fine-tuning.
- 2) Exp. 2: Federated fine-tuning of LLaVA with LoRA.
- 3) Exp. 3: Personalized federated training of CNN3D.

For each experiment, we collect:

- Performance metrics: accuracy, F1 score, and video classification throughput.
- **Communication metrics:** number of rounds, model upload/download size, and adapter payload.
- Environmental metrics: energy consumption (kWh) and carbon emissions (gCO₂e), estimated using the

CodeCarbon library [22]. Measurements are averaged across three independent runs. CO₂ estimates use a grid intensity of 42 gCO₂e/kWh (Paris region).

All training and inference are conducted on NVIDIA A10 GPUs. Default values are used for batch size and frame resolution. Early stopping is applied using a patience of 5 rounds based on validation F1 score.

On Energy Modeling. While this study relies on direct energy measurements, we acknowledge the need for a formal modeling approach that can generalize beyond specific hardware or regional conditions. To this end, we adopt a conceptual framework in which the environmental impact of training or inference consists of two components: (i) operational emissions derived from electricity consumption, grid carbon intensity, and data center efficiency (e.g., Power Usage Effectiveness), and (ii) amortized embodied emissions associated with hardware manufacturing, proportionally allocated based on usage and lifetime.

In this framework, electricity consumption can be approximated using hardware-level specifications such as Thermal Design Power (TDP), utilization rate, duration of workload, and an overhead multiplier to account for system components outside the GPU. This formulation—presented during GT PIA internal working group meetings and inspired by prior sustainable AI methodology, with a formal publication pending—is consistent with models proposed in the sustainable AI literature.

However, for large vision-language models such as LLaVA NeXT Video and Ovis2, no published studies currently provide complete and validated measurements of utilization profiles, workload durations, or embodied emissions. While some isolated benchmarks report inference-level energy draw, the lack of open and detailed lifecycle data precludes full modeling at this stage.

Accordingly, our reported emissions are derived from empirical measurement only. We introduce the above modeling structure as a scientific foundation for future extensions, particularly in scenarios where direct measurement is not feasible or when projecting across heterogeneous deployment infrastructures.

IV. EXPERIMENTS AND RESULTS

A. Exp. 1 : Zero-Shot VLM Inference: Accuracy vs. Efficiency

In this experiment, we evaluate the performance and energy efficiency of large vision-language models (VLMs) used in a zero-shot setting—i.e., without any fine-tuning—on surveillance video classification. This scenario simulates a real-world deployment where pretrained models are directly queried to detect violence using fixed prompts such as: "Analyze the video. Is this a fight scene? Answer yes or no."

1) Evaluation on RLVS.: Table I presents the results on the full RLVS dataset (2000 videos). Ovis2-8B yields the highest accuracy (95.80%) and F1 score (96%) while maintaining a relatively low energy consumption of 457 Wh. Emissions can

be derived from energy using 42 g CO₂/kWh. LLaVA-7B follows closely in accuracy (94.15%) with slightly lower energy usage (328 Wh), highlighting a strong trade-off between precision and frugality. Other models such as Qwen-VL-7B and InternVL3 perform well in accuracy but lack complete environmental metrics. These results demonstrate that off-the-shelf VLMs can achieve high-quality violence detection on RLVS without model adaptation, though model selection remains critical for energy-sensitive deployments.

TABLE I
MEASURED ZERO-SHOT INFERENCE ON RLVS (2,000 VIDEOS);
MEAN + STD OVER 3 RUNS

Model Acc. (%)		F1	Energy (Wh)	
Ovis2-8B	95.80 ± 0.15	0.960 ± 0.004	457 ± 7	
LLaVA-7B	94.15 ± 0.20	94.0 ± 5	328 ± 6	
Qwen-VL-7B	91.00 ± 0.25	91.0 ± 6	616 ± 8	
InternVL3	92.75 ± 0.22	92.0 ± 5	_	
MiniCPM	87.75 ± 0.30	86.0 ± 6	_	
Ovis2-4B	60.75 ± 0.45	_	_	

- indicates missing or unmeasured values.

Energy measured in Watt-hours (Wh).

Emissions can be estimated using 42 g CO₂/kWh = 0.042 g/Wh.

2) Evaluation on RWF-2000 (interpolated).: Since full inference runs on RWF-2000 were not initially available, we extrapolated the results from 400-video benchmarks using linear scaling, which is reasonable given the uniform number of frames per video and deterministic processing pipeline. Table II shows the estimated results for Qwen2.5-Omni, Qwen-VL, and LLaVA-7B. Notably, LLaVA-7B retains strong accuracy (86.75%) with the lowest energy usage (474 Wh) among the tested models. However, performance is weaker compared to RLVS—highlighting that zero-shot VLMs are sensitive to domain shift and dataset-specific content, which is especially relevant in surveillance contexts.

TABLE II EXTRAPOLATED ZERO-SHOT INFERENCE ON RWF-2000, BASED ON 400-video test set, mean \pm std over 3 runs

Model	Acc. (%)	F1 Score (%)	Energy (Wh)
Qwen2.5-Omni-7B	86.75 ± 0.25	86;75 ± 6	2013 ± 30
Qwen2.5-VL-7B	85.75 ± 0.30	85.75 ± 6	872 ± 20
LLaVA-7B	86.75 ± 0.20	86.75 ± 5	474 ± 10

Values scaled from 400-video benchmark using linear extrapolation.

Energy measured in Watt-hours (Wh).

Emissions can be estimated using 42 g CO₂/kWh = 0.042 g/Wh.

3) Combined benchmark on RLVS + RWF (800 videos).: To enable a broader comparison of zero-shot inference performance under consistent conditions, we evaluated four vision-language models—Ovis2-8B, LLaVA-7B, InternVL3, and Qwen-VL—on a shared benchmark of 800 surveillance videos. As shown in Table III, Ovis2-8B achieves the highest accuracy (92.5%), F1 score (92.48%), and ROC AUC (92.5%), while maintaining relatively low inference time and energy usage (1928 s, 172 Wh).

LLaVA-7B performs competitively, especially in terms of F1 (90.62%), but with higher inference time and energy (3215 s, 200 Wh). InternVL3 and Qwen-VL also demonstrate good ROC AUC scores (around 0.885), but with slightly lower

accuracy and, in the case of Qwen-VL, a substantially higher energy footprint (660 Wh) and runtime (5770 s). These results illustrate the trade-offs between predictive performance and efficiency across VLMs, and motivate the selection of LLaVA-7B for further fine-tuning experiments due to its strong balance of performance and adaptability.

TABLE III
ZERO-SHOT INFERENCE: MEASURED PERFORMANCE ON 400 VIDEOS FROM RLVS + 400 VIDEOS FROM RWF-2000 (MEAN \pm STD OVER 3 RUNS)

Model	Accuracy	F1 Score	ROC AUC	Energy	Duration
	(%)	(%)	(%)	(Wh)	(s)
Ovis2-8B	92.5 ± 0.2	92.5 ± 0.4	92.5 ± 0.6	172 ± 4	1928 ± 30
LLaVA-7B	90.4 ± 0.3	90.6 ± 0.5	86.4 ± 0.7	200 ± 5	3215 ± 46
InternVL3	88.5 ± 0.3	88.3 ± 0.4	88.5 ± 0.6	130 ± 4	1745 ± 35
Qwen-VL	88.5 ± 0.3	87.4 ± 0.5	88.5 ± 0.5	660 ± 12	5770 ± 55

4) Takeaways and Summary: Zero-shot VLMs such as Ovis2-8B, LLaVA-7B, InternVL3, and Qwen-VL demonstrate strong performance without fine-tuning. However, accuracy gaps and energy variability highlight the challenge of domain sensitivity—especially in non-IID surveillance environments. To address this, we explore federated fine-tuning and client-side personalization in subsequent experiments.

While Ovis2-8B yields the highest zero-shot accuracy and calibration, we focus our fine-tuning experiments on LLaVA-7B due to its support for LoRA-based adaptation and its mature open-source ecosystem. This makes it a reproducible and scalable candidate for frugal adaptation strategies in federated contexts.

B. Exp. 2: Federated Fine-Tuning of LLaVA with LoRA

In this experiment, we evaluate the benefits of LoRA-based fine-tuning of the LLaVA-7B model under a federated learning setup. We compare its performance against the zero-shot baseline on a heterogeneous, non-IID client distribution, simulating real-world deployment conditions where surveillance content differs across locations.

Although all clients are simulated on a single server, the setup captures key aspects of edge-oriented environments—such as non-IID data, limited participation, and communication-efficient updates via LoRA. This controlled simulation provides a reproducible testbed for evaluating feasibility and trade-offs. Future work will explore deployment on actual edge hardware to validate these findings under real-world constraints.

The full dataset (RWF-2000 + RLVS = 4000 videos) is partitioned across 10 clients: clients 1–5 receive only RWF videos, and clients 6–10 receive only RLVS. A Dirichlet distribution ($\alpha=1$) introduces label imbalance. Each client uses 24-frame clips, and the data is split 80/20 for training and validation (800 videos in total). We simulate partial participation: 5 clients are selected per round (50%) over 20 rounds, using FedAvg. Only LoRA adapter weights are updated and exchanged; the base LLaVA model remains frozen. CodeCarbon estimates are computed under the Île-de-France grid (PUE = 1).

Table IV summarizes the results. Compared to zero-shot inference, federated LoRA fine-tuning significantly improves

calibration (log loss reduced from 0.707 to 0.535) and ROC AUC (85.93 \rightarrow 91.24), while also yielding slight gains in accuracy (+0.25) and decision confidence. Energy consumption for fine-tuning remains reasonable at 570 Wh, compared to 200 Wh for zero-shot inference.

We also tested a stricter setting with only 3 clients per round (30% participation). While this configuration improved ROC AUC further (to 92.59), it yielded slightly lower accuracy and F1 score. We therefore retain the 5-client setup as our main reference.

These results confirm that LoRA-based federated finetuning enhances adaptability and confidence under realistic non-IID settings, without requiring full model updates or high energy budgets.

TABLE IV
LLAVA FINE-TUNED WITH FEDERATED LEARNING(FL) AND LORA
(FEDAVG, 5 CLIENTS/ROUND) VS. ZERO-SHOT ON 800-VIDEO
VALIDATION SET (RWF + RLVS)

Metric	LoRA FL	Zero-Shot
Training Duration (s)	4741	_
Energy (Wh)	570	200
Emissions (g CO ₂ e)	32	8.4
Accuracy (%)	90.87	90.62
F1 Score (%)	90.93	90.84
ROC AUC (%)	91.24	85.93
Log Loss	0.535	0.706

Validation set = 800 videos (20% of RWF + RLVS). Energy measured in Wh. Mean over 3 runs.

C. Exp. 3: Personalized Federated Learning with Lightweight CNN3D

In this experiment, we evaluate a personalized federated learning (PFL) approach based on a compact 3D convolutional neural network (CNN3D) with approximately 65.8M parameters. Unlike LLaVA, which is fine-tuned using LoRA on a frozen backbone, CNN3D uses full end-to-end training but with a decoupled architecture: shared feature extraction layers are aggregated via FedAvg, while client-specific classification heads are trained locally.

The federated setup mirrors Exp. 2: 10 clients (5 with RWF, 5 with RLVS), 20 communication rounds, 5 clients per round. The dataset is split 80/20, and the validation set consists of 800 videos, drawn from the full non-IID distribution. CodeCarbon measurements are computed with consistent assumptions (Îlede-France region, PUE = 1), and training is performed on a single NVIDIA A10 GPU.

Table V reports the results alongside the LLaVA zero-shot and LoRA-adapted models. CNN3D with PFL achieves 90.75% accuracy and 0.9066 F1-score, closely matching LLaVA + LoRA. It also reaches the highest ROC AUC (92.59) and the lowest log loss (.546), indicating strong calibration and generalization under domain shift. Most notably, CNN3D consumes only 240 Wh during training—less than half the energy of LoRA fine-tuning—and emits just $10.1~{\rm g~CO_2e.}$

These results confirm that compact CNN architectures with personalized updates can match the performance of transformer-based models, while consuming significantly

fewer resources—even when accounting for the inference-only cost of zero-shot VLMs like LLaVA (200 Wh). All energy values in this comparison include full training (for CNN3D and LoRA) or full inference (for zero-shot), including model loading where applicable.

TABLE V
FEDERATED PERFORMANCE: CNN3D vs. LLAVA ZERO-SHOT vs.
LLAVA + LORA (20 ROUNDS, 5 CLIENTS/ROUND)

Metric	CNN3D	LLaVA ZS	LLaVA LoRA
Train Time (s)	3795	-	4741
Energy (Wh)	240	200	570
CO ₂ (g)	10.1	8.4	24.0
Acc. (%)	90.75	90.62	90.87
F1 (%)	90.66	90.84	90.93
AUC	92.59	85.93	91.24
LogLoss	5.46	7.07	5.35

ZS = zero-shot. LoRA = LoRA fine-tuned. Acc. = Accuracy, AUC = ROC AUC. CNN3D and LLaVA LoRA are evaluated on a non-IID validation set (800 videos); LLaVA ZS is evaluated on a balanced, centralized 800-video test set

D. Energy and Carbon Footprint Analysis

We now analyze the energy consumption and CO₂ emissions across the three evaluated strategies: zero-shot inference with vision-language models (VLMs), federated LoRA finetuning of LLaVA, and personalized federated training of a lightweight CNN3D model.

From Table I, LLaVA-7B demonstrates efficient zero-shot inference with 200 Wh consumption and 8.4 g CO₂ emissions. By contrast, Qwen2.5-Omni-7B exhibits much higher energy usage (201 Wh) and emissions (113 g). Notably, Ovis2-8B achieves the best accuracy while remaining frugal (457 Wh), confirming that VLM energy profiles vary significantly—underscoring the importance of model selection for sustainable deployment.

In federated training (Table IV), LLaVA fine-tuned with LoRA consumes 570 Wh and emits 24 g CO₂e over 20 rounds across 10 clients. This cost is justified by improved calibration (log loss) and ROC AUC (85.93 \rightarrow 91.24), enabling more confident and reliable predictions in real-world, heterogeneous settings.

By comparison, CNN3D with personalized updates (Table V) delivers nearly equivalent accuracy while consuming only 240 Wh and emitting 10.1 g CO₂e—less than half of LoRA's footprint. It also achieves the best ROC AUC (92.59) and lowest log loss (.546), suggesting that lightweight CNNs with decoupled updates offer both energy efficiency and strong generalization.

We summarize this trade-off in Table VI.

TABLE VI
SUMMARY OF ENERGY AND EMISSIONS ACROSS EVALUATED STRATEGIES

Method	Accuracy (%)	Energy (Wh)	CO ₂ (g)
Zero-shot (LLaVA-7B)	90.62	200	8.4
Fed LoRA (LLaVA-7B)	90.87	570	24.0
PFL CNN3D (65.8M	90.75	240	10.1
params)			

These insights underscore the strategic value of model selection in resource-constrained federated environments. For routine classification tasks, personalized CNNs deliver strong accuracy with minimal energy and emissions. For nuanced inference or multimodal interpretability, VLMs with LoRA remain a viable—though more resource-intensive—choice. This duality supports a hybrid deployment model: use lightweight CNNs by default, and reserve VLMs for selective, high-context queries.

E. Reflections on Lifecycle Energy Attribution and Transfer Overhead

While our evaluation focuses primarily on inference and federated fine-tuning energy consumption, a complete environmental assessment would ideally account for the upstream cost of pretraining large vision-language models (VLMs). Training foundational models such as LLaVA-7B typically entails extensive GPU-hours, substantial memory requirements, and a non-trivial infrastructure footprint. Even when deployed solely for inference, these models carry embedded carbon debt from their pretraining phase. A pragmatic approximation might allocate a fractional share (e.g., 1/1000th) of the original training emissions per deployment use case, as suggested by prior studies [14], [15]. While such amortized lifecycle costs are not included in our current figures, they would further accentuate the relative efficiency of lightweight, task-specific models such as CNN3D in repeated or distributed settings.

In addition, centralized training paradigms—particularly with high-resolution, sensitive video data—introduce often-overlooked network energy costs. Uploading raw surveillance footage from edge devices to a central server for model training incurs significant bandwidth and power demands, especially when multiplied across thousands of geographically distributed clients. In contrast, federated learning transmits only compact model updates (e.g., LoRA adapter weights or personalized CNN parameters), dramatically reducing transfer overhead. The net benefit depends on network topology, data compression, and training frequency, but early estimates suggest that federated approaches can offer both privacy protection and lower communication-related emissions.

These considerations highlight that environmental sustainability in AI extends beyond algorithmic efficiency—it requires system-level thinking that includes model reuse, data logistics, and the full lifecycle impact of deployment architectures.

V. DISCUSSION

Our comparison between LoRA-tuned LLaVA and a personalized CNN3D model highlights a critical trade-off for sustainable and privacy-aware AI deployment. Despite a nearly two orders of magnitude difference in parameter count—7 billion for LLaVA versus 65.8M for CNN3D—both models achieve comparable performance, surpassing 90% accuracy and yielding similar F1 scores. Notably, CNN3D slightly outperforms in terms of ROC AUC and log loss, suggesting better calibration under non-IID conditions.

The environmental contrast is even more striking: CNN3D requires approximately **half the energy** (240 Wh vs. 570 Wh)

and emits less than half the $CO_{2}e$ (10.1 g vs. 24 g) compared to LoRA-based fine-tuning. These results underscore the effectiveness of parameter-frugal personalization for energy-conscious federated learning.

While LoRA-tuned vision-language models (VLMs) offer advantages in multimodal reasoning and interpretability, their higher computational cost makes them more suitable for selective use—e.g., in situations requiring nuanced scene understanding, long-range dependencies, or human-centric queries. In contrast, lightweight CNNs serve as reliable and efficient backbones for routine classification tasks in video surveillance.

These results motivate—but do not prescribe—a strategic hybrid deployment pattern, where lightweight CNNs handle routine inference by default, and transformer-based VLMs are selectively triggered in complex or ambiguous cases.

Future directions include developing context-aware model selection mechanisms, optimizing client-side inference latency, and extending evaluation to richer tasks such as video captioning, anomaly explanation, or real-time alert prioritization.

VI. CONCLUSION

This study had two main objectives: (i) assess whether large vision-language models (VLMs) can operate efficiently in zero-shot or federated contexts, and (ii) evaluate whether lightweight, personalized models can offer comparable performance with lower environmental cost.

We conducted a systematic comparison between LoRA-tuned LLaVA-7B and a compact CNN3D model trained via personalized federated learning. Both were tested under realistic non-IID conditions, with explicit tracking of energy consumption and CO_2 emissions.

Key findings include:

- Both models exceeded 90% accuracy on validation sets;
- CNN3D slightly outperformed LLaVA in ROC AUC and log loss, while using less than half the energy;
- LoRA-tuned VLMs offered richer multimodal reasoning and interpretability;
- Zero-shot inference remains viable but energy and domain-sensitivity trade-offs must be considered.

These results highlight the value of hybrid deployment strategies: using efficient CNNs for routine inference, and selectively activating VLMs for complex or descriptive scenarios. They also reinforce the importance of frugal design beyond training—considering inference efficiency and lifecycle emissions.

Future work under the **DIVA** initiative will explore adaptive model selection, multimodal fusion (e.g., audio-text), and privacy-aware optimization on edge devices. We also aim to expand evaluation frameworks to include structured outputs and full lifecycle analysis.

REFERENCES

[1] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, "Flamingo: A

- visual language model for few-shot learning," in Advances in Neural Information Processing Systems (NeurIPS), 2022.
- [2] Z. Li, X. Wu, H. Du, H. Nghiem, and G. Shi, "Benchmark evaluations, applications, and challenges of large vision-language models: A survey," Research Square, preprint, 2025. DOI: 10.32388/GXR68Q.
- [3] D. P. Nguyen, J. P. Muñoz, and A. Jannesari, "FLoRA: Enhancing vision-language models with parameter-efficient federated learning," arXiv preprint arXiv:2404.15182, 2024.
- [4] M. Jiang, T. Jung, R. Karl, and T. Zhao, "Federated dynamic graph neural networks with secure aggregation for video-based distributed surveillance," ACM Trans. Intell. Syst. Technol., vol. 13, no. 4, pp. 1–24, 2022
- [5] A. Singh, Z. Khan, S. Kp, and S. H., "Understanding federated learning and its application in video anomaly detection," *Int. J. Eng. Appl. Sci. Technol.*, vol. 7, no. 1, pp. 10–15, 2022.
- [6] S. Abdulrahman, H. Tout, H. Ould-Slimane, A. Mourad, and C. Talhi, "A survey on federated learning: The journey from centralized to distributed on-site learning and beyond," *IEEE Internet Things J.*, vol. 8, no. 7, pp. 5476–5497, 2021.
- [7] P. Bellavista, L. Foschini, and A. Mora, "Decentralised learning in federated deployment environments," ACM Comput. Surv., vol. 54, no. 9, pp. 1–38, 2021.
- [8] H. Vyas and P. M. Torrens, "Tackling 'leaky' facial and object recognition operations over federated learning," in *Proc. 4th Interdisciplinary Conf. on Electrics and Computer (INTCEC)*, 2024.
- [9] L. G. R. Ribeiro, M. Léonardon, G. Muller, V. Fresse, and M. Arzel, "FLoCoRA: Federated learning compression with low-rank adaptation," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2024.
- [10] J. Bian, L. Wang, L. Zhang, and J. Xu, "LoRA-FAIR: Federated LoRA fine-tuning with aggregation and initialization refinement," arXiv preprint arXiv:2403.09922, 2024.
- [11] L. Yi, H. Yu, G. Wang, and X. Liu, "FedLoRA: Model-heterogeneous personalized federated learning with LoRA tuning," arXiv preprint arXiv:2307.04579, 2023.
- [12] J. Qi, Z. Luan, S. Huang, C. Fung, H. Yang, and F. Sha, "FDLoRA: Personalized federated learning of large language models via dual LoRA tuning," arXiv preprint arXiv:2403.13478, 2024.
- [13] Y. Su, N. Yan, and Y. Deng, "Federated LLMs fine-tuned with adaptive importance-aware LoRA," arXiv preprint arXiv:2403.11380, 2024.
- [14] A. Lacoste, A. Luccioni, V. Schmidt, and T. Dandres, "Quantifying the carbon emissions of machine learning," arXiv preprint arXiv:1910.09700, 2019.
- [15] P. Henderson, J. Hu, J. Romoff, E. Brunskill, D. Jurafsky, and J. Pineau, "Towards the systematic reporting of the energy and carbon footprints of machine learning," *J. Mach. Learn. Res.*, vol. 21, no. 248, pp. 1–43, 2020.
- [16] M. Kassir, S. Haidar, and A. Yaacoub, "Exploring personalized federated learning architectures for violence detection in surveillance videos," arXiv preprint arXiv:2504.00857, 2025.
- [17] Q. Pajon, S. Serre, H. Wissocq, L. Rabaud, S. Haidar, and A. Yaa-coub, "Balancing accuracy and training time in federated learning for violence detection in surveillance videos: A study of neural network architectures," *J. Comput. Sci. Technol.*, vol. 39, no. 5, pp. 1029–1039, Sep. 2024. DOI: 10.1007/s11390-024-3702-7.
- [18] M. Cheng, K. Cai, and M. Li, "RWF-2000: An open large scale video database for violence detection," in *Proc. 25th Int. Conf. on Pattern Recognition (ICPR)*, Milan, Italy, Jan. 2021, pp. 4183–4190. DOI: 10.1109/ICPR48806.2021.9412502.
- [19] M. Soliman, M. Kamal, M. Nashed, Y. Mostafa, B. Chawky, and D. Khattab, "Violence recognition from videos using deep learning techniques," in *Proc. 9th Int. Conf. on Intelligent Computing and Information Systems (ICICIS)*, Cairo, Egypt, Dec. 2019, pp. 79–84.
- [20] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, pp. 1273–1282, 2017.
- [21] E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," arXiv preprint arXiv:2106.09685, 2021.
- [22] M. Anthony, F. Luccioni, V. Ligozat, and A. Lacoste, "Carbontracker: Tracking and predicting the carbon footprint of training deep learning models," arXiv preprint arXiv:2007.03051, 2020.