Federated Learning for Video Violence Detection: Complementary Roles of Lightweight CNNs and Vision-Language Models for Energy-Efficient Use

Sébastien Thuau
esieaLab, ETIS Laboratory
ESIEA, University of CY Cergy
Paris, France
thuau@et.esiea.fr

Siba Haidar
esieaLab
ESIEA
Paris, France
siba.haidar@esiea.fr

Rachid Chelouah
ETIS Laboratory, CNR1S, UMR8051
University of CY Cergy
Paris, France
rc@cy-tech.fr

***Abstract*—Deep learning–based video surveillance increasingly demands privacy-preserving architectures with low computational and environmental overhead. Federated learning preserves privacy but deploying large Vision-Language Models (VLMs) introduces major energy and sustainability challenges. We compare three strategies for federated violence detection under realistic non-IID splits on RWF-2000 and RLVS datasets: zero-shot pretrained VLM inference, LoRA-based fine-tuning of LLaVA-NeXT-Video-7B, and personalized federated learning of a 65.8 M-parameter 3D CNN. All methods exceed 90% accuracy in binary violence detection. The 3D CNN achieves superior calibration (ROC AUC 92.59%) at roughly half the energy cost (240 Wh vs. 570 Wh) of federated LoRA, while VLMs provide richer multimodal reasoning. Hierarchical category grouping—based on semantic similarity and class exclusion—boosts VLM multiclass accuracy from 65.31% to 81% on the UCF-Crime dataset.**

To our knowledge, this is the first comparative simulation study of LoRA-tuned VLMs and personalized CNNs for federated violence detection, with explicit energy and CO_2e quantification. Our results inform hybrid deployment strategies that default to efficient CNNs for routine inference and selectively engage VLMs for complex contextual reasoning.

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

I. Introduction

Deep learning–based video surveillance systems are increasingly deployed to detect and analyze violent incidents in public spaces, yet centralized training introduces privacy risks by transferring sensitive footage off-device and incurs substantial computational and environmental overhead.

Vision-Language Models (VLMs) enhance multimodal reasoning through natural language interfaces but carry billions of parameters, challenging energy efficiency and carbon footprint constraints in federated learning environments, which—while preserving privacy—do not inherently address the resource demands of such architectures.

Aligned with the goals of the DIVA (Decentralized Intelligence for Visual Awareness) initiative, we investigate whether large-scale VLMs can be fine-tuned sustainably under non-IID federated conditions for violence detection and compare their energy consumption and emissions against lightweight 3D CNNs with competitive performance.

Model naming convention.: After first mention, we use the following abbreviations: *LLaVA-NeXT-Video-7B* (**LLaVA**); *Qwen2.5-VL-7B-Instruct* (**QwenVL**); *Qwen2.5-Omni-7B* (**QwenOmni**); *Ovis2 (8B)* (**Ovis-8B**); *Ovis2 (4B)* (**Ovis-4B**); *InternVL3-8B* (**InternVL-8B**); *MiniCPM-o-2.6* (**MiniCPM**). We also refer to *Personalized Federated Learning* as (**PFL**) and to the personalized 3D convolutional network as (**3D CNN**).

To this end, we assess zero-shot inference of a VLM baseline, federated LoRA fine-tuning of LLaVA, and personalized federated learning of a compact 3D CNN for binary classification on RWF-2000 and RLVS datasets with realistic non-IID splits. Our key contributions include the first comparative study of LoRA-tuned VLMs versus personalized CNNs for federated violence detection, a systematic quantification of energy use, communication overhead, and CO_2e emissions, evaluation under distributed non-IID conditions, and practical insights for hybrid deployment strategies that balance accuracy, privacy, and sustainability.

II. Related Work

A. Privacy-Preserving Architectures for Video Surveillance and Energy-Efficient Multimodal AI Models

Federated learning (FL) mitigates privacy risks inherent to centralized video analytics by keeping raw footage on-device and exchanging only model updates or task-specific features. Compared to server-side ingestion of video streams, FL removes large-scale data transfers and reduces exposure of personally identifiable information. Prior systems for video-based distributed surveillance train local models at the edge and synchronize via secure aggregation, preventing the server from inspecting client updates [1], [2]. Surveyed frameworks further layer formal protections—differential privacy, homomorphic encryption, and multi-party computation—to strengthen confidentiality guarantees [3], [4]. Complementary "semantic-first" pipelines discard raw frames after on-device processing and

transmit only compact, AI-extracted representations, aligning with frugal deployment principles [5].

On the efficiency front, parameter-efficient fine-tuning (PEFT) enables low-cost personalization in FL. Low-Rank Adaptation (LoRA) factorizes weight updates into low-rank matrices, substantially reducing compute and communication. Recent federated LoRA frameworks report large efficiency gains: aggregation-aware and quantization-enhanced designs shrink update sizes with minor accuracy loss, while server-side correction mitigates aggregation bias without added bandwidth [6], [7]. For multimodal settings, FLORA applies rank-2 LoRA adapters to CLIP ViT-L/14 and demonstrates significant speedups and memory savings over full fine-tuning across diverse vision datasets [8]. Additional variants address heterogeneity via shared homogeneous adapters, dual (global/personalized) adapters, and adaptive-rank tuning for constrained clients [9], [10], [11]. Collectively, these methods fall into (i) aggregation-enhancing strategies (e.g., stacking or corrected aggregation) and (ii) adapter-level optimizations (quantization, personalization, adaptive rank).

Despite this progress, the literature seldom quantifies the cost–benefit trade-offs of federated fine-tuning for vision–language models (VLMs) versus lightweight CNNs under realistic non-IID surveillance data. Moreover, to our knowledge, no prior work evaluates LoRA-based federated VLM fine-tuning for violence detection under practical surveillance conditions or benchmarks its cost-effectiveness against lightweight CNNs under matched non-IID partitions—questions we address empirically.

B. Environmental Impact Assessment

At surveillance scale, the cumulative energy footprint of model training and adaptation across thousands of cameras is non-negligible, yet empirical transparency on consumption and emissions remains limited. Emerging policy and standardization efforts underscore the need for rigorous environmental accounting throughout the AI lifecycle. In the EU, the AI Act introduces obligations touching energy transparency and resource use for higher-risk systems, while French initiatives promote methodological baselines for "frugal" AI (e.g., AFNOR SPEC 2314) and institutional support via ADEME and related programs. Building on this context, we quantify energy and carbon for centralized vs. federated fine-tuning of VLMs, and compare against frugal CNN baselines under non-IID deployments. Our results provide evidence to guide sustainable surveillance model selection and adaptation.

III. Methodology

A. Binary Classification

1) Experimental Design and Data Configuration: We build a 4,000-clip corpus by concatenating RWF–2000 and RLVS (2,000 samples each; balanced violence/non–violence) [12], [13]. A stratified 80/20 split defines train/validation. To induce cross–silo heterogeneity, data are partitioned over 10 clients by domain: clients 1–5 receive only RWF–2000, clients 6–10 only RLVS. Additional non–IID skew is introduced via Dirichlet sampling with concentration α=1 at the client level.

2) Distributed Learning Configuration: All FL runs are simulated (no real devices). We train with FedAvg for 20 communication rounds [14], emulating 10 clients (each: 1 GPU, 4 CPU cores). Each round, we sample 50% of clients (5/10) to model partial participation due to bandwidth/availability.

3) Vision–Language Model Adaptation: We apply LoRA [15] to **LLaVA** (*LLaVA–NeXT–Video–7B*) [16], [17]. The visual encoder is CLIP ViT–L/14 [18], operating on 24–frame clips. Prompts follow: ``Analyze the video. Is this a fight scene? Answer yes or no.'' Adapters use rank r=16, scaling α=32, with 4–bit quantization; only adapter weights are communicated. Adapter modules are inserted in attention blocks; base weights remain frozen.

4) Comparative Baseline Architecture: As a personalized baseline, we use a ∼65.8M–param 3D–CNN with parameter decoupling [27]: a globally aggregated spatiotemporal backbone (shared features) and client–local classification heads (specialization), enabling personalization while preserving cross–domain transfer.

5) Performance Assessment Protocol: We evaluate under three regimes: (i) zero–shot VLMs (**LLaVA**, **QwenVL**, **QwenOmni**, **Ovis-8B**, **InternVL-8B**, **MiniCPM**), (ii) federated LoRA fine–tuning, and (iii) personalized CNN training. Metrics: accuracy and F1 Score (reported as %). Environmental impact is computed with CodeCarbon [19], converting metered energy (kWh) to gCO_2e using a Paris–region factor of 56 gCO_2e/kWh (assumption). Hardware: NVIDIA A10; results are averaged over three seeds.

6) Theoretical Basis for Environmental Impact Modeling: We propose a dual framework for modeling the carbon footprint of machine learning workflows—distinguishing runtime emissions from hardware-related manufacturing costs—offering a scalable foundation when empirical measurement is limited or impractical.

B. Multiclass Classification

To probe beyond binary decisions, we evaluate VLM comprehension on multiclass anomaly recognition using UCF–Crime, a large–scale surveillance benchmark covering 13 anomaly categories *plus a Normal class* [20], [21]. As a zero–shot baseline, we run inference on 20% of the data with three VLMs reused from the binary study (**QwenVL**, **Ovis-8B**, **LLaVA**), then extend inference to the full dataset to validate error patterns. Confusion analysis reveals systematic overlaps among semantically close classes. We therefore adopt a hierarchical grouping on the full set to reduce ambiguity while preserving operational relevance:

- **Destruction**: Arson, Explosion
- **Property Crime**: Burglary, Robbery, Shoplifting, Stealing, Vandalism
- **Violence**: Assault, Fighting

RoadAccidents and *Normal Video* are retained as distinct categories; *Abuse*, *Arrest*, and *Shooting* are excluded. The

inclusion/exclusion rationale and downstream effects on accuracy and calibration are detailed in Section IV.

IV. Experiments and Results

A. Binary Classification

1) Exp. 1.1: Zero-Shot VLM Inference: Accuracy vs. Efficiency: We assessed the performance–efficiency trade-offs of pretrained vision-language models on binary violence classification without fine-tuning. For readability and consistency with the above convention, we use **LLaVA**, **QwenVL**, **QwenOmni**, **Ovis-8B**, **InternVL-8B**, and **MiniCPM** throughout the rest of the paper. Performance evaluation on RLVS (2,000 videos) revealed that **Ovis-8B** [22] achieved optimal accuracy (95.80%) and F1 Score (96.0%) while maintaining reasonable energy consumption (457 Wh). **LLaVA** [16] demonstrated competitive performance (94.15% accuracy) with lower energy requirements (328 Wh), as shown in Table I.

TABLE I
Zero-shot VLM performance on RLVS dataset (2000 videos), mean ± std over 3 runs

Model	Acc. (%)	F1 Score(%)	Energy (Wh)
Ovis-8B [22]	**95.80 ± 0.15**	**96.0 ± 0.4**	**457 ± 7**
LLaVA [16]	94.15 ± 0.20	94.0 ± 5.0	328 ± 6
QwenVL [23]	91.00 ± 0.25	91.0 ± 6.0	616 ± 8
InternVL-8B [24]	92.75 ± 0.22	92.0 ± 5.0	–
MiniCPM [25]	87.75 ± 0.30	86.0 ± 6.0	–
Ovis-4B [22]	60.75 ± 0.45	–	–

– indicates missing values. Energy in Wh.
Emissions: 56 g CO_2e/kWh = 0.056 g/Wh.

A comprehensive benchmark on 800 videos (400 RLVS + 400 RWF-2000) confirmed model performance variations across architectures. Table II demonstrates **Ovis-8B**'s superior performance (92.5% accuracy, 92.5% ROC AUC) with efficient resource utilization (172 Wh, 1928 s inference time).

TABLE II
Comprehensive zero-shot evaluation on 800 surveillance videos (mean ± std over 3 runs).

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

ROC AUC scores were computed only when the model produced consistent confidence outputs across classes; otherwise, metrics were omitted or approximated from logits or ranking confidence.

Carbon emissions were estimated using regional grid intensity (56 g CO_2e/kWh) with CodeCarbon [19] monitoring. Results demonstrated significant variance in energy efficiency across architectures, with LLaVA selected for subsequent federated experiments due to its balanced performance–efficiency profile and LoRA compatibility.

2) Exp. 1.2: Federated Fine-Tuning of LLaVA with LoRA: We implemented parameter-efficient federated adaptation using LoRA [15] on LLaVA under non-IID conditions. The experimental setup partitioned 4,000 videos across 10 clients with domain-based heterogeneity: clients 1-5 received RWF-2000 content, clients 6-10 processed RLVS data. Label imbalance was induced by sampling from a Dirichlet distribution with a concentration parameter of one. All federated experiments employed partial participation, with 5 out of 10 clients selected at random in each round to reflect realistic connectivity and bandwidth limitations.

The federated optimization employed FedAvg [14] with partial participation (5/10 clients per round) over 20 communication epochs. LoRA adapters utilized rank-16 configuration with scaling factor $\alpha = 32$, implemented through 4-bit quantization via the Flower federated learning framework [26]. Differential privacy mechanisms were integrated using fixed clipping (norm=1) to ensure convergence stability.

Table III presents the comparative analysis between federated LoRA fine-tuning and zero-shot baseline. Federated adaptation significantly improved model calibration (log loss: 0.706 $\rightarrow$ 0.535) and discriminative performance (ROC AUC: 85.93% $\rightarrow$ 91.24%) while maintaining competitive accuracy.

TABLE III
Federated LoRA adaptation vs. zero-shot baseline on 800-video validation set

Metric	LoRA FL	Zero-Shot
Training Duration (s)	4741	–
Energy (Wh)	570	200
Emissions (g CO_2e)	32.0	11.2
Accuracy (%)	90.87	90.62
F1 Score (%)	90.93	90.84
ROC AUC (%)	91.24	85.93
Log loss	0.535	0.706

Validation: 800 videos (20% of RWF + RLVS). Mean over 3 runs. FL = Federated Learning.

Training consumed 570 Wh over 4,741 seconds, generating 32.0 g CO_2e emissions. The efficiency gains from transmitting only 4-bit adapter weights demonstrated the viability of communication-efficient federated adaptation for resource-constrained surveillance environments.

3) Exp. 1.3: Personalized Federated Learning with Lightweight 3D CNN: We evaluated a parameter-decoupled 3D convolutional architecture (65.8M parameters) using personalized federated learning (PFL) [27]. The model employed shared spatiotemporal feature extraction layers (globally aggregated) and client-specific classification heads (locally optimized), balancing generalization with domain adaptation.

The federated configuration mirrored previous experiments: 10 clients with domain-based partitioning, 20 communication rounds, and 50% participation rate. Table IV presents comparative performance across the three evaluated approaches.

TABLE IV
Performance comparison: 3D CNN vs. LLaVA variants (20 rounds, 5 clients/round)

Metric	3D CNN PFL	LLaVA ZS	LLaVA LoRA
Train Time (s)	3795	–	4741
Energy (Wh)	240	200	570
CO_2e (g)	13.4	11.2	31.9
Acc. (%)	90.75	90.62	90.87
F1 Score(%)	90.66	90.84	90.93
AUC (%)	**92.59**	85.93	91.24
Log loss	**0.546**	0.707	0.535

ZS = zero-shot, PFL = Personalized FL. 3D CNN and LoRA: non-IID validation (800 videos). ZS: balanced centralized test set (800 videos).

Energy efficiency analysis demonstrated superior resource utilization: 240 Wh consumption and 13.4 g CO_2e emissions during complete federated training. The 3D CNN achieved the highest ROC AUC (92.59%) and the lowest log loss (0.546) among evaluated approaches.

B. Multiclass Classification with Vision-Language Models

1) Zero-shot Baseline Evaluation: To evaluate the performance of vision-language models (VLMs) on multiclass crime detection, we conducted inference on the complete UCF-Crime dataset using zero-shot classification. This approach allows us to assess the models' inherent understanding of criminal activities without task-specific fine-tuning.

a) Results: Table V presents the overall performance metrics for all three models on the whole dataset of UCF-Crime, with the initial 14 classes as defined in the dataset.

TABLE V
ZERO-SHOT BASELINE EVALUATION RESULTS (MEAN ± STD OVER 3 RUNS)

Model	Accuracy (%)	F1 Score (%)	Energy (Wh)	CO_2e (g)
Ovis-8B	65.31	60.43	1149.8	64.4
LLaVA	47.77	38.01	385.2	21.6
QwenVL	51.26	48.35	597.6	33.5

These findings indicate marked disparities in model performance, with **Ovis-8B** surpassing all others on every metric. Accordingly, subsequent results are presented for **Ovis-8B**. With the best model achieving only 65.31% accuracy, we investigated the factors underlying the VLMs' difficulties.

b) Per-Class Analysis: Figure 1 illustrates the precision and recall performance for each category across **Ovis-8B**. The analysis reveals substantial variation in detection accuracy across different crime types, with certain categories proving more challenging for automated classification. Furthermore, the recall for "Normal Video" is exceptionally high because the models tend to label inputs as "Normal Video" when uncertain.

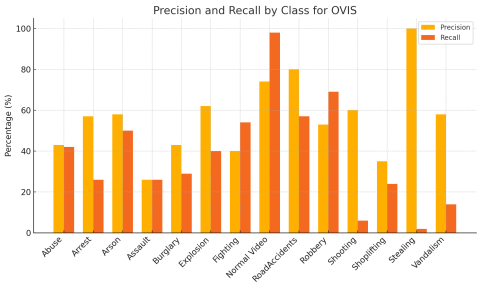

Fig. 1. Precision and recall scores by class for Ovis-8B on the full UCF-Crime dataset. Significant variability across classes highlights the difficulty of fine-grained classification and supports the motivation for semantic grouping.

Figure 2 shows the normalized confusion matrix with "Normal Video" excluded to focus on anomaly classification.

This confusion matrix (Fig. 2) highlights pronounced misclassifications among semantically adjacent categories by **Ovis-8B**. The three most significant confusions are:

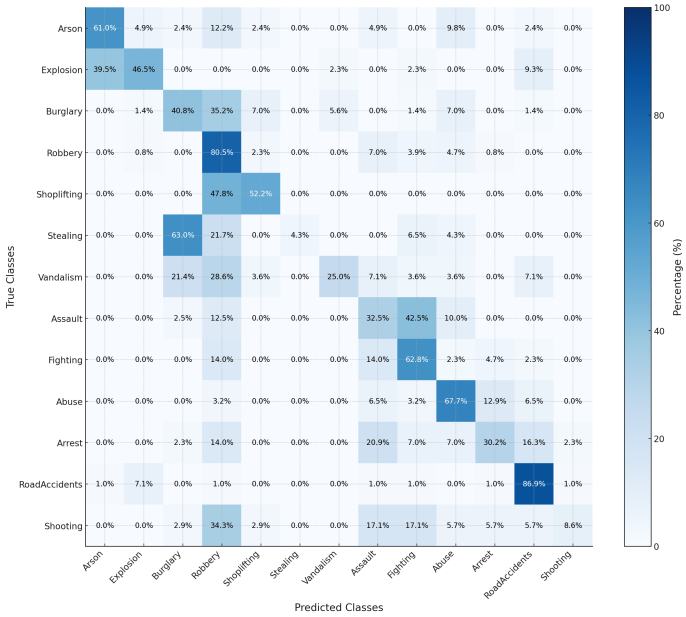

Fig. 2. Normalized confusion matrix for zero-shot VLM predictions on the full UCF-Crime dataset, *excluding* the "Normal Video" class. Highlights widespread misclassification between semantically adjacent classes and shows the tendency of VLMs to default to "Normal" when uncertain.

1) *Arson* vs. *Explosion*, since explosions frequently involve fire;
2) *Burglary*, *Robbery*, *Shoplifting*, *Stealing*, and *Vandalism*, all concerning offenses against property, where the model fails to distinguish subtle differences among property crimes;
3) *Assault* vs. *Fighting*, both depicting interpersonal violence.

Moreover, the *Arrest* and *Shooting* classes exhibit pervasive confusion with other anomalies and were therefore excluded from further experiments. The *Abuse* class was excluded owing to the subtle and ambiguous nature of its instances.

2) Refined Classification Strategy: The revised classification strategy now focuses on aggregated categories as described in Section III-B. We performed inference again on the full dataset, excluding the 150 videos from the omitted classes, and instructed the model to predict only the remaining categories. The same prompt structure was reused with grouped categories replacing the original 14 classes.

The results, shown in Table VI, exhibit a marked improvement in performance, as expected given the reduced number of target classes.

TABLE VI
ZERO-SHOT BASELINE WITH REFINED CLASSIFICATION (MEAN ± STD OVER 3 RUNS)

Model	Accuracy (%)	F1 Score (%)	Energy (Wh)	CO_2e (g)
Ovis-8B	81.0	80.0	985.9	55.2
LLaVA	50.0	41.0	346.7	19.4
QwenVL	70.0	69.0	931.5	52.2

Figure 3 shows the confusion matrix for **Ovis-8B**'s category inference. Overall misclassification rates are reduced relative to the previous matrix, except within the *Property Crime* category, which still overlaps with *Destruction* and *Violence*.

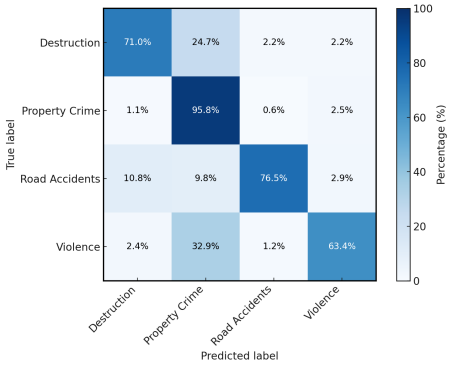

Fig. 3. Normalized confusion matrix after semantic grouping of UCF-Crime classes into four categories. While overall classification performance improves, some residual confusion persists, especially between *Violence* and *Property Crime*.

V. Conclusion

We present a comprehensive head-to-head comparison of federated LoRA-tuned vision–language models (VLMs) and personalized lightweight 3D CNNs for video violence detection under realistic non-IID settings. All approaches exceed 90% binary accuracy; the CNN achieves stronger calibration (ROC AUC: 92.59%, log loss: 0.546) while using roughly half the energy of federated LoRA (240 Wh vs. 570 Wh). For multiclass, hierarchical grouping improves VLM accuracy (65.31% $\rightarrow$ 81%), and LoRA-based adaptation enhances VLM calibration under non-IID clients. Note that the CNN baseline was evaluated only for binary detection, whereas multiclass analysis was conducted exclusively with VLMs.

These findings crystallize an efficiency–capability trade-off: lightweight CNNs suit continuous, low-cost screening on constrained nodes, while VLMs should be selectively invoked for high-context reasoning and natural-language incident descriptions. We therefore recommend a hybrid cascade with always-on CNN screening and on-demand VLM escalation for complex or ambiguous events.

Our environmental accounting—via direct energy metering and emissions estimation—indicates that such tiered operation reduces energy and CO_2e without degrading situational awareness. Future work focuses on template-driven VLM reporting, uncertainty/novelty-conditioned model selection, multiclass 3D CNN, tighter PEFT, and lifecycle-aware assessment within **DIVA**.

References

[1] 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.

[2] 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.

[3] S. AbdulRahman, H. Tout, H. Ould-Slimane, A. Mourad, C. Talhi, and M. Guizani, "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, Apr. 2021. doi:10.1109/JIOT.2020.3030072.

[4] P. Bellavista, L. Foschini, and A. Mora, "Decentralised Learning in Federated Deployment Environments," *ACM Comput. Surv.*, vol. 54, no. 9, pp. 1–38, 2021.

[5] H. Vyas and P. M. Torrens, "Tackling 'Leaky' Facial and Object Recognition Operations over Federated Learning," in *Proc. INTCEC*, 2024.

[6] L. G. R. Ribeiro, M. Léonardon, G. Muller, V. Fresse, and M. Arzel, "FLoCoRA: Federated Learning Compression with Low-Rank Adaptation," in *Proc. EUSIPCO*, 2024.

[7] J. Bian, L. Wang, L. Zhang, and J. Xu, "LoRA-FAIR: Federated LoRA Fine-Tuning with Aggregation and Initialization Refinement," arXiv:2403.09922, 2024.

[8] D. P. Nguyen, J. P. Muñoz, and A. Jannesari, "FLoRA: Enhancing Vision-Language Models with Parameter-Efficient Federated Learning," arXiv:2404.15182, 2024.

[9] L. Yi, H. Yu, G. Wang, and X. Liu, "FedLoRA: Model-Heterogeneous Personalized Federated Learning with LoRA Tuning," arXiv:2307.04579, 2023.

[10] 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:2403.13478, 2024.

[11] Y. Su, N. Yan, and Y. Deng, "Federated LLMs Fine-Tuned with Adaptive Importance-Aware LoRA," arXiv:2403.11380, 2024.

[12] 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)*, 2021, pp. 4183–4190. doi:10.1109/ICPR48806.2021.9412502.

[13] 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.

[14] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks From Decentralized Data," in *Proc. AISTATS*, 2017, pp. 1273–1282.

[15] 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:2106.09685, 2021.

[16] Y. Zhang *et al.*, "LLaVA-NeXT: A Strong Zero-Shot Video Understanding Model," [Online]. Apr. 2024. Available: https://llava-vl.github.io/blog/2024-04-30-llava-next-video/. [Accessed: 24-Sep-2025].

[17] H. Liu *et al.*, "LLaVA-NeXT: Improved Reasoning, OCR, and World Knowledge," [Online]. Jan. 2024. Available: https://llava-vl.github.io/blog/2024-01-30-llava-next/. [Accessed: 24-Sep-2025].

[18] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," in *Proc. ICML*, PMLR 139, 2021, pp. 8748–8763.

[19] CodeCarbon contributors, "CodeCarbon: Estimate and Track Carbon Emissions from Machine Learning Computing," Software, 2021–2025. [Online]. Available: https://mlco2.github.io/codecarbon/. [Accessed: 24-Sep-2025].

[20] W. Sultani, C. Chen, and M. Shah, "Real-World Anomaly Detection in Surveillance Videos," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6479–6488.

[21] CRCV, University of Central Florida, "Real-World Anomaly Detection in Surveillance Videos (UCF-Crime)," [Online]. Available: https://www.crcv.ucf.edu/projects/real-world/. [Accessed: 24-Sep-2025].

[22] S. Lu *et al.*, "Ovis: Structural Embedding Alignment for Multimodal Large Language Model," arXiv:2405.20797, 2024.

[23] Qwen Team, "Qwen2.5-VL," [Online]. Jan. 2025. Available: https://qwenlm.github.io/blog/qwen2.5-vl/. [Accessed: 24-Sep-2025].

[24] J. Zhu *et al.*, "InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models," arXiv:2504.10479, 2025.

[25] Y. Yao *et al.*, "MiniCPM-V: A GPT-4V-Level MLLM on Your Phone," arXiv:2408.01800, 2024.

[26] D. J. Beutel *et al.*, "Flower: A Friendly Federated Learning Research Framework," arXiv:2007.14390, 2020.

[27] A. Z. Tan, H. Yu, L. Cui, and Q. Yang, "Towards Personalized Federated Learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 12, pp. 9587–9603, Dec. 2023. doi:10.1109/TNNLS.2022.3160699.