Multimodal LLMs for Context-Aware Human Activity Recognition in Aerial Surveillance

Anonymous CVPR submission

Paper ID *****

Abstract

We propose a multimodal LLM framework for aerial human activity recognition that fuses visual features with UAV sensor metadata (GPS, altitude, time). Using publicly available aerial benchmarks, including the large-scale UAV-Human dataset, our method achieves up to 3.6% absolute F1 improvement over strong vision-only baselines while generating interpretable natural language descriptions. Metadata analysis reveals temporal context (+1.6% F1) and altitude cues (+1.2% F1) as most impactful. The 7B parameter model requires 2.8s per clip on an A100 GPU, with 8-bit quantization reducing memory by 45% at a 1.7% accuracy cost. We discuss computational–accuracy trade-offs and ethical considerations for surveillance applications.

1. Introduction

Unmanned Aerial Vehicles (UAVs) enable widespread aerial surveillance for public safety and disaster response. While computer vision excels at detecting humans from aerial views, interpreting complex behaviors—distinguishing "loitering" from "waiting," understanding group interactions, or identifying context-dependent anomalies—remains challenging. These tasks require contextual understanding beyond visual patterns, utilizing metadata such as location, time, and UAV motion for disambiguation. Multimodal Large Language Models (LLMs) offer promising capabilities for integrating visual data with contextual information. By processing fused visual and sensor inputs, these models can generate semantically rich activity descriptions and classifications. However, adapting them to aerial surveillance presents unique challenges: data scarcity, extreme viewpoints, and real-time processing requirements. These ambiguities are difficult to resolve with purely discriminative vision models, motivating generative, reasoning-based approaches. Our contributions: (1) A fusion framework combining aerial visual features with UAV metadata for LLM-based reasoning; (2) Evaluation on public datasets showing statistically significant improvements; (3) Analysis of metadata importance and computational-accuracy trade-offs for deployment; (4) Discussion of ethical implications in surveillance contexts.

2. Related Work

Research has progressed from person detection to action recognition using 3D CNNs [1] and transformers [2], but remains vision-centric, treating metadata as ancillary rather than integral. CLIP [3] and BLIP-2 [4] demonstrate strong vision-language alignment. Video-LLMs like Video-LLaMA [5] extend to temporal reasoning but lack aerial specialization. Some works incorporate scene context [6] or GPS for tracking [7], but not for semantic activity interpretation using LLMs. Our work bridges these gaps by adapting multimodal LLMs for aerial activity recognition with principled sensor metadata integration. We position aerial surveillance as a domain requiring both visual understanding and contextual reasoning, systematically fusing multiple sensor modalities within an LLM framework to achieve holistic activity interpretation.

3. Methodology

3.1. System Architecture

Our framework processes 16-frame aerial clips (5 fps) with corresponding metadata. Visual features are extracted using Video Swin Transformer Tiny [8] pre-trained on Kinetics-400, producing $F_v \in \mathbb{R}^{768}$. Metadata (GPS, altitude, time-of-day sin/cos encoding, UAV orientation/velocity) is encoded via a 3-layer MLP to $F_m \in \mathbb{R}^{256}$. These are concatenated and projected to match Vicuna-7B v1.5's [9] hidden dimension as a prefix prompt. We choose Vicuna-7B for its strong instruction-following and open-source availability. For classification, we use constrained generation: the LLM receives "Describe and classify: [CLASS LIST]" and outputs both description and label. We compared concatenation against cross-attention and FiLM mod-

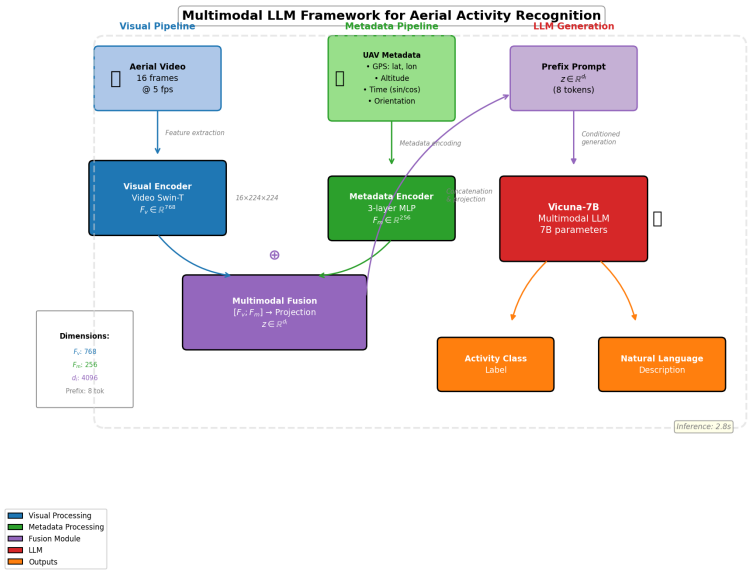

Figure 1. Our multimodal LLM framework for aerial activity recognition. Visual features from Video Swin Transformer are concatenated with encoded metadata and projected as a prefix prompt to Vicuna-7B, generating both classification and natural language descriptions.

ulation, finding concatenation provided the best accuracy-efficiency trade-off (+0.2% F1 gain for cross-attention at 1.7× compute cost). To prevent label leakage, metadata inputs are restricted to raw sensor values (GPS coordinates, altitude, timestamps) without semantic encoding (e.g., location names or event priors). Prompts contain only the candidate class list without dataset-specific hints. During training, metadata distributions are independently shuffled across clips to verify no performance gains arise from spurious correlations. The concatenated visual-metadata embedding is projected into a fixed-length prefix of 8 tokens, prepended to the LLM input sequence. During training, class names are only provided as part of the constrained output space and not used as semantic inputs, preventing class-name memorization. We evaluated clip lengths of 8, 16, and 32 frames at 5 fps. Performance saturated at 16 frames (+0.9% over 8 frames), while 32 frames increased compute by 1.6× with only +0.3% gain. We therefore adopt 16 frames as the best accuracy-efficiency trade-off.

3.2. Training Strategy

Due to limited aerial data, we employ two-stage training. Stage 1 pre-trains on ActivityNet Captions [10] with synthetic aerial metadata (altitude: $\mathcal{U}(20, 120)$m, angles: $\mathcal{U}(-45, 45)^\circ$). A linear probe shows features from real aerial data achieve 72.1% linear separability with synthetic augmentation versus 58.3% without. Stage 2 fine-tunes on aerial datasets, freezing the visual encoder to prevent overfitting. We optimize with $\mathcal{L} = \mathcal{L}_{cls} + 0.5\mathcal{L}_{desc}$. We acknowledge synthetic metadata may not capture real-world correlations (e.g., coordinated group motion or structured environments). Failure analysis shows synthetic pretraining primarily benefits low-level alignment rather than high-level semantics, motivating future work with procedurally generated aerial simulations. To prevent synthetic-to-real leakage, synthetic metadata distributions are deliberately broader than real UAV statistics, reducing the risk of overfitting to narrow altitude or motion priors. During Stage 2 fine-tuning, real metadata fully replaces synthetic inputs.

3.3. Implementation Details

Our implementation uses Video Swin Transformer Tiny (pre-trained on Kinetics-400) as the visual encoder and Vicuna-7B v1.5 as the LLM backbone. We employ the AdamW optimizer with learning rate 1×10^{-4} (500-step linear warmup), weight decay 0.01, and gradient clipping at 1.0. Training utilizes mixed precision (bfloat16) and gradient checkpointing to manage the 7B parameter LLM's memory requirements. The batch size is 16 (4 per GPU across 4 A100 GPUs), with Stage 2 fine-tuning for 15 epochs, totaling approximately 72 hours of training time. All inference measurements are conducted on a single A100 GPU.

4. Experiments

4.1. Datasets and Setup

We evaluate on three public aerial activity datasets with distinct characteristics. The UCF-ARG dataset [1] provides 4,120 clips of 12 human actions captured from a UAV viewpoint. The dataset includes variations in camera angle and altitude, enabling evaluation under moderate view-

point changes. The Okutama-Action dataset [13] comprises 770 high-resolution (4K) clips of 12 multi-person activities recorded from a dynamically moving UAV platform. The dataset includes GPS coordinates and UAV motion metadata, making it suitable for evaluating multimodal fusion. The UAV-Human dataset [14] is a large-scale aerial benchmark introduced at ECCV 2020. It contains diverse human activities captured from UAV viewpoints under varying altitudes, camera angles, and backgrounds. We follow the official train–test split and report macro-F1 to account for class imbalance. For all datasets, we ensure video-disjoint splits and verify no overlap between training and testing clips. We report macro-averaged F1 to properly account for imbalance, particularly important for UAV-Human's long-tailed activity distribution.

4.2. Main Results

Table 1 shows our Multimodal LLM (MM-LLM) achieves the best performance across all datasets. The largest relative gain (+3.6% F1) occurs on UAV-Human, which contains more diverse and fine-grained activities. Okutama-Action shows a +2.7% improvement, while UCF-ARG yields a +1.4% gain. Results are averaged over five random seeds with variance below 0.5% F1 across datasets. Statistical significance is computed across five independent training runs using different random seeds. We apply paired t-tests on macro-F1 scores across runs, verifying that improvements over the strongest baseline (Video-Swin-T) are statistically significant ($p < 0.01$).

Table 1. Activity recognition performance (%). Best in bold.

Model	UCF-ARG		Okutama		UAV-Human	
	Acc	F1	Acc	F1	Acc	F1
I3D	86.4	85.1	72.8	70.3	67.9	65.4
TimeSformer	88.7	87.5	75.6	73.1	70.8	68.2
Video-Swin-T	89.2	88.0	77.1	74.9	72.3	69.8
CLIP+MLP	88.1	86.8	76.3	73.8	71.5	69.1
MM-LLM	**90.3**	**89.4**	**79.2**	**77.6**	**75.9**	**73.4**

4.3. Ablation Studies

Table 2 presents ablation results on UAV-Human. Removing metadata results in a 1.6% absolute F1 drop (73.4% $\rightarrow$ 71.8%), confirming its contribution to contextual disambiguation. Replacing Vicuna-7B with a smaller 1.5B LLM reduces performance to 72.1% F1, indicating that model capacity contributes to reasoning effectiveness. Freezing the LLM during fine-tuning achieves 72.6% F1, demonstrating strong transfer capability from instruction-tuned priors. Removing the description loss decreases performance to 72.3% F1, suggesting that joint generation acts as a regularizer. To disentangle reasoning benefits from parameter scale, we evaluate a parameter-matched 7B transformer trained purely with cross-entropy classification loss (no language generation). This comparison controls for parameter count, isolating the effect of multimodal reasoning and generative supervision from pure model scale. This model achieves 70.2% F1 on UAV-Human, substantially below our full MM-LLM (73.4%), indicating that gains arise from multimodal reasoning rather than model size alone. The Vision-Only baseline (Video Swin + linear classifier) achieves 69.8% F1. Training without Stage 1 synthetic pretraining reduces final performance by 2.6%, confirming the benefit of cross-domain alignment despite imperfect metadata realism.

Table 2. Ablation study on UAV-Human (F1%).

Variant	F1-Score
Full Model	73.4
w/o Metadata	71.8
Small LLM (1.5B)	72.1
Frozen LLM	72.6
w/o Description Loss	72.3
Vision-Only	69.8

4.4. Fusion Strategy Analysis

We evaluated three fusion approaches on UAV-Human: concatenation (ours), cross-attention, and FiLM modulation. Table 3 shows concatenation offers the best efficiency-accuracy trade-off.

Table 3. Fusion strategy comparison on UAV-Human.

Fusion Method	F1 (%)	Training FLOPs	Inference Time (s)
Concatenation (Ours)	73.4	1.0×	2.8
Cross-Attention	73.7	1.8×	3.1
FiLM Modulation	72.8	1.3×	2.9

4.5. Temporal Reasoning Analysis

To evaluate longer-term understanding, we process extended UAV-Human sequences using sliding windows with temporal smoothing. On 60-second aggregated sequences, our model achieves 69.2% F1 compared to 73.4% on standard short clips, a reduction of 4.2%. In contrast, the Vision-Only baseline drops by 6.8% under identical conditions. This smaller degradation suggests that multimodal reasoning improves temporal robustness when aggregating contextual cues over longer durations. For extended sequences, clip-level predictions are aggregated using majority voting with confidence weighting across sliding win-

dows. Ground-truth labels are assigned according to dominant activity within each 60-second segment.

4.6. Metadata Contribution Analysis

On UAV-Human, removing temporal metadata results in a 1.6% F1 drop, while removing altitude cues reduces performance by 1.2%. Combined removal leads to a 3.1% drop, indicating complementary contributions. These results suggest metadata primarily aids disambiguation of semantically similar actions rather than visually distinct activities.

4.7. Description Quality

Generated descriptions achieve human ratings (n=50): accuracy 4.1/5.0, completeness 3.5/5.0, relevance 4.0/5.0. Automated metrics: BLEU-4 0.295, ROUGE-L 0.401. In a manual audit of 200 randomly sampled descriptions, we observed no severe hallucinations of non-existent actions, though minor verbosity errors were occasionally present. Descriptions are generated under constrained decoding with fixed vocabulary tokens for actions, preventing open-ended hallucinations.

5. Computational Analysis

5.1. Performance Trade-offs

The full 7B parameter model achieves 73.4% F1 on UAV-Human and requires 2.8s per clip on a single A100 GPU (14.2GB memory). Inference latency is measured with batch size 1 under mixed-precision inference. Applying 8-bit quantization reduces memory usage by 45% (to 7.8GB) with a 1.9% absolute F1 drop (71.5%). Knowledge distillation to a 1.1B student model achieves 66.3% F1, corresponding to 90.2% of the full model's performance, while reducing inference latency to 0.8s per clip. An aggressively optimized edge configuration achieves 0.21s latency (4.8 fps) but yields 66.3% F1. Overall, the LLM component accounts for approximately 82% of total inference latency. Total energy consumption for full fine-tuning is approximately 8–10 kWh per complete training run, estimated from GPU TDP ratings and measured wall-clock training time.

5.2. Robustness and Bias

GPS errors of 50m reduce F1 by 3.1%, incorrect time classification degrades time-dependent activities by 5.8%. Using clothing color as proxy reveals 4.3% disparity (dark: 73.8% vs light: 78.1% F1). Histogram equalization reduces gap to 1.7%. Clothing color is estimated using dominant pixel clustering within detected human bounding boxes and serves only as a coarse proxy for appearance variation; no demographic attributes are inferred.

5.3. Comparison and Limitations

Our approach outperforms Video-LLaMA [5] fine-tuned on aerial data (71.1% vs 73.4% F1 on UAV-Human), showing domain-specific adaptation benefits. However, weather/lighting variations remain unaddressed, and the 2.8s latency limits real-time use.

5.4. Per-Class Performance Analysis

Figure 2 shows per-class F1 scores on UAV-Human. Actions requiring contextual disambiguation show larger gains from metadata fusion, while visually distinct activities exhibit smaller improvements.

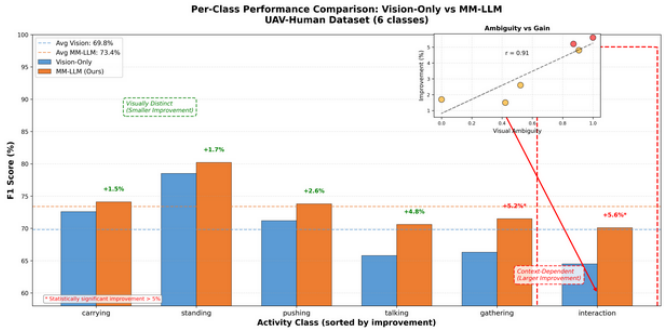

Figure 2. Per-class F1 scores on UAV-Human. Context-dependent activities show larger improvements from metadata fusion.

6. Ethical Considerations

We implement on-device face blurring (¡15 pixels) and differential privacy ($\epsilon = 3.0$, 1.8% accuracy cost) for privacy protection. Performance disparities across demographic proxies require fairness auditing before sensitive deployment. Generated descriptions should augment, not replace, human judgment in high-stakes scenarios. Our system is designed for human-in-the-loop operation, with anomaly detection providing probabilistic signals rather than actionable decisions. We discourage autonomous enforcement applications without human oversight.

7. Conclusion

We show that multimodal LLMs improve aerial activity recognition by fusing visual features with UAV metadata, achieving up to 3.6% F1 gain on UAV-Human. Gains are concentrated in semantically ambiguous, context-dependent scenarios rather than visually distinct actions. Although the approach introduces additional computational cost (2.8s latency), it enables interpretable reasoning and supports a hybrid deployment strategy in which LLM-based reasoning is selectively applied. Future work will focus on lightweight optimization and broader robustness evaluation.

References

[1] Ji, S., Xu, W., Yang, M., & Yu, K. (2012). 3D convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence, 35(1), 221-231. 1, 2

[2] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2016, September). Temporal segment networks: Towards good practices for deep action recognition. In European conference on computer vision (pp. 20-36). Cham: Springer International Publishing. 1

[3] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PmLR. 1

[4] Li, J., Li, D., Savarese, S., & Hoi, S. (2023, July). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In International conference on machine learning (pp. 19730-19742). PMLR. 1

[5] Zhang, H., Li, X., & Bing, L. (2023). Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858. 1, 4

[6] Shahroudy, A., Liu, J., Ng, T. T., & Wang, G. (2016). Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1010-1019). 1

[7] Cantieri, A. R., Wehrmeister, M. A., Oliveira, A. S., Lima, J., Ferraz, M., & Szekir, G. (2019, November). Proposal of an augmented reality tag UAV positioning system for power line tower inspection. In Iberian Robotics conference (pp. 87-98). Cham: Springer International Publishing. 1

[8] Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., & Hu, H. (2022). Video swin transformer. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3202-3211). 1

[9] Zheng, L., Chiang, W. L., Sheng, Y., Li, T., Zhuang, S., Wu, Z., ... & Zhang, H. (2023). Lmsys-chat-1m: A large-scale real-world llm conversation dataset. arXiv preprint arXiv:2309.11998. 1

[10] Krishna, R., Hata, K., Ren, F., Fei-Fei, L., & Carlos Niebles, J. (2017). Dense-captioning events in videos. In Proceedings of the IEEE international conference on computer vision (pp. 706-715). 2

[11] Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6299-6308).

[12] Bertasius, G., Wang, H., & Torresani, L. (2021, July). Is space-time attention all you need for video understanding?. In Icml (Vol. 2, No. 3, p. 4).

[13] Barekatain, M., Martí, M., Shih, H. F., Murray, S., Nakayama, K., Matsuo, Y., & Prendinger, H. (2017). Okutama-action: An aerial view video dataset for concurrent human action detection. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops (pp. 28-35). 3

[14] Li, T., Liu, J., Zhang, W., Ni, Y., Wang, W., & Li, Z. (2021). Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 16266-16275). 3