# Generalized VLM-Based In-Car Violence Detection Leveraging Synthetic Data

SeifEldin Zeidan
*Media Engineering and Technology*
*German University in Cairo*
Cairo, Egypt
seifeldin.zeidan@gmail.com

Shashank Mishra
*Augmented Vision*
*German Research Center for Artificial Intelligence (DFKI)*
Kaiserslautern, Germany
shashank.Mishra@dfki.de

Jason Rambach
*Augmented Vision*
*German Research Center for Artificial Intelligence (DFKI)*
Kaiserslautern, Germany
jason_raphael.rambach@dfki.de

Mohammed A.-Megeed Salem
*Media Engineering and Technology*
*German University in Cairo*
Cairo, Egypt
mohammed.salem@guc.edu.eg

***Abstract*—As autonomous vehicles move toward full driverless operation, ensuring passenger safety through in-cabin monitoring is critical, yet violence detection remains challenging due to the scarcity of real-world datasets. Traditional approaches rely on supervised CNN-based architectures, which often fail to generalize across different environments. In this paper, we evaluate the transition from task-specific models to large-scale Video-Language Models (VLMs), specifically InternVideo2-Chat, for identifying aggressive interactions. We demonstrate that while the Temporal Shift Module (TSM) achieves high accuracy on seen distributions, it suffers a huge drop in performance on out-of-distribution data. In contrast, Low-Rank Adaptation (LoRA) fine-tuned VLM matches state-of-the-art performance (98.7% accuracy) while maintaining generalization. Furthermore, we investigate the potential of AI-generated content's ability to replace real-world captured datasets by building a synthetic in-car violence dataset using the open source Wan2.2 diffusion model. The results demonstrate that fine-tuning solely on generated video data leads to a substantial improvement of approximately 29% points in recall over the zero-shot baseline, while maintaining a high accuracy of 90.5% on real-world test data, suggesting that generative models can effectively mitigate data scarcity for rare and sensitive events. Our findings highlight that semantic-driven VLMs offer a more scalable and robust solution for the future of shared autonomous mobility security.**

***Index Terms*—VLMs, Violence, Diffusion, In-vehicle, LoRA, SAVs.**

## I. Introduction

Autonomous driving is rapidly changing how we travel. As shared autonomous vehicles (SAVs) replace human drivers, the responsibility for passenger safety shifts entirely to the vehicle's onboard systems. This makes in-cabin monitoring essential for detecting unusual behavior and preventing conflicts or violence between passengers.

While major progress has been made in areas like autonomous driving, pedestrian detection, and lane tracking [1]–[3], far less attention has been given to what happens inside the vehicle, especially detecting violent interactions. With the growth of shared mobility and ride-hailing services, developing systems that can automatically detect such behavior is increasingly important.

Recent advances in computer vision have significantly improved action recognition performance on large-scale benchmarks such as Kinetics [4] and Something-Something [5]. However, real-world in-vehicle violence detection presents unique challenges that are not fully addressed by these datasets. Existing research on violence detection has primarily focused on surveillance environments such as public spaces, streets, or indoor security footage.

Traditional approaches to video action recognition in such domains rely heavily on supervised convolutional architectures. For example, 3D CNN-based architectures have shown effectiveness in modeling spatiotemporal features for action and violence detection tasks. Similarly, the Temporal Shift Module (TSM) [6] has demonstrated strong performance in video understanding by efficiently modeling temporal dependencies within 2D convolutional backbones. However, these supervised models often suffer from limited generalization capabilities; although they perform well within the same environment or dataset they are trained on, their performance typically degrades when evaluated in unseen or different environments.

At the same time, advances in Large Language Models (LLMs) and Video-Language Models (VLMs) are changing how these problems are approached. Instead of relying on task-specific, supervised models, the focus is shifting toward large-scale multimodal foundation models that can handle a wide range of tasks. For example, InternVideo2 [7] presents a scalable video foundation model that achieves state-of-the-art results on action recognition benchmarks. When paired with an LLM in InternVideo2-Chat, it also supports open-ended reasoning and classification through natural language prompts, making it a more flexible alternative to traditional supervised methods.

Recent progress in AI-driven video generation especially with diffusion-based text-to-video models has made it possible to generate temporally coherent video clips using structured text prompts. As a result, they present a promising alternative to traditional data collection methods, which are often expensive and hard to collect. This can help with the creation of synthetic datasets that can supplement or even replace real-world data. The remainder of this paper is organized as follows: Section II reviews related work. Section III describes the proposed methodology. Section IV shows the experimental results. Section V covers the discussion and finally, Section VI concludes the paper.

## II. RELATED WORK

Early research on in-car monitoring mainly focused on the driver, aiming to detect behaviors such as distraction and drowsiness [8]–[10]. These systems were designed under the assumption that the driver remains the central agent responsible for vehicle safety and passenger well-being. However, in fully autonomous settings, occupants are exclusively passengers, and this transition introduces new challenges, as there is no longer a human supervisor to intervene in critical situations.

There is a shortage of high-quality violence detection datasets due to privacy concerns and the fact that real-world violent events are rare. Most existing research relies on datasets from public surveillance, such as the UCF-Crime dataset [11], which focuses on anomalies in security footage. Similarly, foundational benchmarks like Hockey Fights [12] and Violent-Flows [13] were developed to identify aggression in sports and crowded public areas. However, these datasets don't really capture what happens in tight, confined spaces like inside a car. As a result, there's still a major gap in research when it comes to detecting violence in these kinds of environments.

Recent studies have begun addressing passenger-focused safety monitoring within vehicle cabins. In particular, Rodrigues et al. [14] introduced the MoLa InCar dataset, a multimodal in-car violent action dataset comprising 20 action scenarios divided into violent and non-violent sequences recorded inside a vehicle. This dataset was created to fill that gap, focusing specifically on capturing instances of violence within the confined space of a car. Later on, in [15], the MoLa InCar dataset was used to benchmark several state-of-the-art video action recognition architectures, including I3D [4], R(2+1)D [16], SlowFast [17], TSN [18], and TSM [6]. Their results demonstrated that TSM [6], pretrained on the Kinetics dataset [4], achieved the highest performance among the evaluated models, reaching 98.27% accuracy on RGB videos. These findings highlight the effectiveness of temporal modeling techniques for violence detection in constrained vehicular environments.

The introduction of Large Language Models (LLMs), has shifted the field from task-specific architectures to general-purpose learners based on the Transformer framework [19]. This progress has recently expanded into the multimodal domain through the development of Vision-Language Models (VLMs). By utilizing strategies like contrastive learning in models such as CLIP [20] or generative pre-training in BLIP [21], VLMs successfully align visual features with semantic text embeddings. This alignment allows models to understand not just what an image contains, but also the nuanced context and descriptions associated with visual data.

The InternVideo series, developed by Shanghai AI Lab, represents a progressive evolution of large-scale video foundation models aimed at unified multimodal and spatiotemporal understanding. The original InternVideo model [22] introduced a general-purpose video representation framework that combines video–text contrastive learning to jointly learn temporal and semantic representations. InternVideo2 [7] further scales model capacity and improves multimodal alignment through a structured three-stage training strategy. The first stage employs masked video modeling to capture fine-grained temporal dependencies. The second stage introduces video–text contrastive learning with a dedicated text encoder to enhance cross-modal alignment. In the final stage, a Q-Former module is used to project video features into a shared embedding space compatible with large language models such as Mistral or InternLM. Low-Rank Adaptation (LoRA) is applied to fine-tune the language backbone efficiently.

Diffusion models have emerged as a dominant framework for generative modeling by learning to progressively denoise random noise into structured samples through a learned reverse diffusion process. The Wan family [23] represents an open-source text-to-video diffusion framework designed for high-quality short-form video generation. In particular, Wan2.2-T2V-A14B is a text-to-video model capable of synthesizing short video clips from structured textual prompts.

Synthetic data generation is generally divided into two streams: graphics engine-based and generative-based methods. The former uses 3D models to render training samples via predefined rules, a technique widely adopted in object detection [24], [25], optical flow estimation [26], and autonomous driving [27]. More recently, the rise of AI-Generated Content has shifted focus toward generative models. While early approaches utilized GANs [28], current research favors diffusion models for their superior realism and controllability. Notably, Li et al. [29] demonstrated that text-to-video diffusion can generate synthetic data that improves the zero-shot performance of Vision-Language Models (VLMs), especially for action recognition tasks where real-world data is limited.

In our work, we integrate recent breakthroughs in generative AI and multimodal learning to address the data scarcity and environmental constraints of violence detection. Our contributions are as follows:

- We evaluate the Vision-Language Model (VLM), InternVideo2-Chat, in both zero-shot and LoRA fine-tuned configurations, comparing its performance against the TSM (Temporal Shift Module) baseline pre-trained on Kinetics for identifying in-vehicle violence.
- We build a synthetic, multi-purpose in-car violence dataset using the Wan2.2 model, systematically generating both violent and non-violent actions inspired by Mola's action set.

- We compare the generalization capabilities between the semantic-driven VLM and the traditional CNN-based TSM architecture.

## III. METHODOLOGY

### A. Dataset

The primary dataset used in this work is the MoLa dataset, an in-car violence detection dataset designed to capture interactions between passengers inside a vehicle. The dataset consists of 20 action scenarios, including 12 violent and 8 non-violent scenarios, which together form 58 distinct actions. Each scenario contains 64 video samples, resulting in a total of 1280 videos recorded at 30 frames per second. In this work, only the RGB modality is used for training and evaluation.
The dataset follows the same segment-wise annotation and evaluation protocol adopted by the original authors [15]. Violent videos typically begin with a non-violent segment before the violent interaction occurs. Therefore, each violent video is divided into two segments: a non-violent segment and a violent segment, based on the annotated start time of the violent event. The average duration is approximately 7 seconds for violent segments and 12 seconds for non-violent segments. During evaluation, the model is assessed segment-wise, treating each segment as an independent sample.
To ensure consistency and direct comparability with the original work, we also adopted the same subject-based dataset split. It also ensures that subjects appearing in the training set do not appear in the validation or test sets. The dataset was divided into training (50%, subjects P1–P8), validation (25%, subjects P9–P12), and testing (25%, subjects P13–P16).
To perform both the segment division and the subject-based split, we used the MOLAnnotate toolkit provided by the authors. Scenario descriptions were then used to create training prompts, replacing binary labels with textual annotations.

### B. VLM: InternVideo2-Chat

To investigate whether large Video–Language Models (VLMs) can match traditional supervised architectures for in-car violence detection, we adopt InternVideo2-Chat-HD as our VLM.

*1) Zero-Shot:* We first evaluate the model in a zero-shot setting to measure its baseline performance without task-specific training. Since zero-shot performance is highly dependent on prompt design, we evaluated different prompts on the validation set.

*2) LoRA Fine-Tuning:* During fine-tuning of the InternVideo2-Chat model as shown in fig. 1, we freeze both the video encoder and the Q-Former, and only update the projection layer and the large language model (LLM) components. To efficiently adapt the LLM, we employ Low-Rank Adaptation (LoRA) fine-tuning. Specifically, we use the same LoRA adapter configuration introduced by the InternVideo2 authors when integrating the video encoder with the LLM, ensuring architectural consistency with the original model design and keeping the rest of the model frozen.

In total, we only fine-tune around 48 million parameters, corresponding to approximately 0.577% of the full model. This parameter scale is comparable to the 24 million trainable parameters used by the TSM baseline, placing the trainable parameter counts within the same order of magnitude.

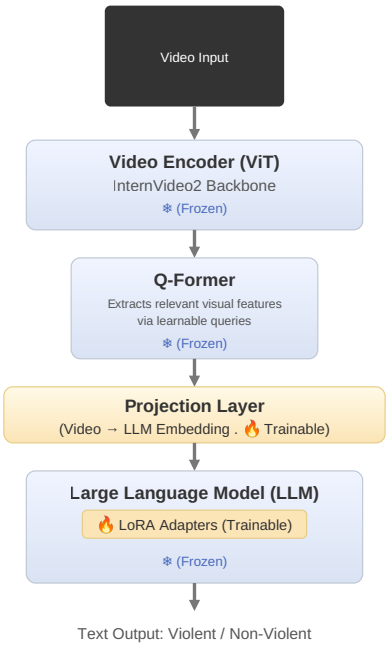


Fig. 1. Trainable Layers for InternVideo2-Chat

### C. Generated Synthetic Dataset

To investigate whether synthetic data could replace real-world datasets, we generated a general purpose in-car violence dataset using the open-source Wan2.2 text-to-video diffusion model. Specifically, we used the Wan2.2-T2V-A14B model to generate a total of 673 video clips, 4 seconds each, including 341 non-violent and 332 violent samples.

The generated dataset includes 29 actions inspired by the MoLa dataset. To ensure diversity and consistency in the generated samples, each generation prompt was constructed using three components:

- Character description
- Action description
- Camera view specification

The character description defines the gender combination of the two passengers (e.g., male–male, male–female, female–female). The action description specifies the interaction taking place inside the vehicle, and for each action we generated approximately four variations. The action description variations were generated using ChatGPT-5 in order to introduce linguistic diversity while preserving the semantic meaning of the actions. The camera view specification describes the position and orientation of the in-car surveillance camera.

For each video, a prompt was generated by randomly combining one character description, one action variation, and

one camera configuration using a script. An example prompt used for video generation is shown below:

"In the back seat of a car, two women sit side by side; the woman on the left turns sideways in the seat, face tense with anger, and throws a rapid series of short punches toward the woman on the right's face. The woman on the right jerks back with each hit, eyes wide and lips parted in shock, one arm raised to shield. A static camera focuses on the entire backseat of the car aligned with the center of the backseat and the two passengers, using a wide shot lens where both passengers are fully visible."

Fig. 2. Examples From The Generated Dataset

Examples of the generated clips are shown in Fig. 2. All generated clips were manually reviewed to verify the label correctness and that no major visual or temporal artifacts were present. Clips with incorrect, ambiguous, or distorted content were rejected. This generated dataset is available upon request.

To evaluate the potential of synthetic data as an alternative to real-world data, we fine-tune the InternVideo2-Chat-HD model using only this generated dataset, without incorporating any samples from the MoLa training set. The training prompts are obtained by extracting a subset of the action description portion of the video generation prompts and by using ffmpeg, we added random noise to the videos to better reflect real-world data characteristics.

During training, the MoLa validation set is used solely for validation and best checkpoint selection, without contributing to any parameter updates. After fine-tuning, the model is evaluated on the MoLa test set to measure its performance on real-world in-car scenarios.

The obtained results are compared against the models trained directly on the MoLa dataset. This experimental setup enables us to determine whether synthetic video data has reached a level that can effectively replace real-world video data for fine-tuning, potentially allowing future systems to rely on AI-generated videos instead of costly, time-consuming and unreliable real-world data collection, especially for anomaly events like violence.

### D. Generalization Comparison

As hypothesized, supervised CNN-based models such as TSM are expected to achieve strong performance on data drawn from the same distribution as their training set, but may have limited generalization when evaluated on different environments. To investigate this, we compare the generalization capabilities of a CNN-based model (TSM) with a Vision-Language Model (InternVideo2-chat).

First, we reproduce the TSM baseline results on the MoLa dataset then both models (trained on MoLa only) will be evaluated on the generated dataset, which serves as an out-of-distribution test environment. This setup enables a direct comparison of how well each model generalizes to unseen environments beyond the one observed during training.

### E. Evaluation Metrics

We evaluate model performance using standard classification metrics, including accuracy, mean class accuracy (MCA), precision, recall, and F1-score. These metrics are defined as:

$$\begin{aligned} \text{Accuracy} &= \frac{TP+TN}{TP+TN+FP+FN}, \\ \text{MCA} &= \frac{1}{2}\left(\frac{TP}{TP+FN}+\frac{TN}{TN+FP}\right), \\ \text{Precision} &= \frac{TP}{TP+FP}, \qquad \text{Recall} = \frac{TP}{TP+FN}, \\ \text{F1-score} &= \frac{2\,\text{Precision}\,\text{Recall}}{\text{Precision}+\text{Recall}}. \end{aligned} \tag{1}$$

where $TP$, $TN$, $FP$, and $FN$ stands for true positives, true negatives, false positives, and false negatives, respectively.

### F. Experimental Setup

All experiments were conducted on an NVIDIA H200 GPU within an Ubuntu environment. During the fine-tuning of the InternVideo2-Chat-HD model, a batch size of 8 was used with gradient accumulation steps set to 8. The learning rate was set to $2 \times 10^{-5}$. During inference, 16 uniformly sampled frames were used. InternVideo2-Chat-HD consumed approximately 22.3 GB of GPU memory under the evaluated configuration. Furthermore, do_sample=False was set in the LLM to prevent randomness in the generated outputs.

For video generation using the Wan 2.2 model, samples were generated using either the UniPC or DPM++ samplers, with the number of inference steps ranging between 60 and 100. A different random seed was used for each generation to benefit from diversity across the dataset.

## IV. Results

### A. Zero-shot

After several prompts were evaluated on the validation set, we found the best-performing prompt was:

"You are a surveillance violence detection assistant. Does this video show any act, threat, or attempt of physical harm; forceful contact; or weapon-related aggression (including displaying, brandishing, or using an object as a weapon), even if

brief, partially visible, or without clear injury? Answer with only: Yes or No, then provide one brief sentence explaining your decision."

As shown in Table I, the zero-shot model achieves 79.3% accuracy and an F1-score of 64.7% on the test set. Although the overall accuracy is relatively high, the recall is significantly lower (51.9%), indicating that the model frequently misses violent events.

### B. LoRA Fine-Tuning

After fine-tuning on the MoLa dataset using the same prompt as the zero-shot, the InternVideo2-Chat model achieves 98.7% accuracy and an F1-score of 98.2%, closely matching the performance of the TSM baseline.

TABLE I
PERFORMANCE COMPARISON ON MOLA DATASET

| Model | Train-Dataset | Evaluation Metrics (%) | | | | |
|---|---|---|---|---|---|---|
| | | *Accuracy* | *MCA* | *Precision* | *Recall* | *F1-score* |
| TSM | MoLa | 98.5 | 98.6 | 96.9 | 98.9 | 97.9 |
| IV2-Chat-HD | Zero-shot | 79.3 | 73.5 | 86.0 | 51.9 | 64.7 |
| | MoLa | 98.7 | 98.6 | 97.9 | 98.4 | 98.2 |
| | GenDataset | 90.5 | 88.4 | 92.7 | 80.4 | 86.1 |

[a]MoLa: MoLa Dataset, Zero-shot: No Training, GenDataset: Generated synthetic dataset.

### C. Generated Dataset

When fine-tuned using the synthetic generated dataset instead of MoLa, the model achieves 90.5% accuracy and an F1-score of 86.1%.

### D. Generalization Comparison

Table II presents the generalization performance of both models when trained on the MoLa dataset and evaluated on the generated dataset. A significant performance gap can be observed between the two approaches. The InternVideo2-Chat-HD model achieves strong results across all metrics, with an accuracy of 91.1% and an F1-score of 91.6%, indicating its ability to generalize effectively to unseen, out-of-distribution data. In contrast, the TSM model shows a substantial drop in performance, achieving only 52.7% accuracy and an F1-score of 8.1%. Although TSM attains a precision of 100%, this is accompanied by an extremely low recall of 4.2%, suggesting that the model predicts very few positive (violent) samples and fails to detect the majority of them.

TABLE II
GENERALIZATION PERFORMANCE ON GENERATED DATASET (MODELS TRAINED ON MOLA)

| Model | Evaluation Metrics (%) | | | | |
|---|---|---|---|---|---|
| | *Accuracy* | *MCA* | *Precision* | *Recall* | *F1-score* |
| TSM (MoLa $\rightarrow$ GenDataset) | 52.7 | 52.1 | 100 | 4.2 | 8.1 |
| IV2-Chat-HD (MoLa $\rightarrow$ GenDataset) | 91.1 | 91.2 | 85.2 | 99.1 | 91.6 |

[a]MoLa $\rightarrow$ GenDataset: Trained on MoLa and tested on generated dataset.

## V. DISCUSSION

The zero-shot results of InternVideo2-Chat-HD highlight the strength of large-scale pre-training, achieving close to 80% accuracy without any training on the MoLa dataset. However, the recall remains relatively low (51.9%), indicating that while the model understands the general concept of violence, it struggles to detect more subtle interactions inside a vehicle. This suggests that although Video-Language Models (VLMs) offer a strong starting point, task-specific fine-tuning is still required for reliable performance in safety-critical applications.

The experiments with the generated dataset show that synthetic data has the potential to be a practical alternative to expensive and time-consuming real-world dataset collection. Fine-tuning on the Wan2.2 generated dataset improved total accuracy from 79.3% to 90.5% and the recall from 51.9% (zero-shot) to 80.4%, demonstrating that synthetic data can significantly enhance model performance when real data is limited.

However, some limitations were observed in the generated videos. The violent actions produced by the Wan2.2 model were generally more subtle and less pronounced compared to those in the real dataset. This reduced expressiveness can make the distinction between violent and non-violent behavior less obvious. In addition, the model struggled with complex physical interactions, such as accurately modeling thrown object trajectories, and often required highly detailed prompts to correctly represent passenger reactions. These observations indicate that current open-source video generation models still have limitations in capturing realistic physical dynamics and interaction details. Nevertheless, as text-to-video models continue to improve, particularly in terms of physical realism and motion consistency, the gap between synthetic and real data is expected to narrow. In the long term, this could enable training and evaluation to rely primarily on synthetic data, reducing both the cost and time required for data collection while also addressing privacy concerns.

After LoRA fine-tuning, IV2-Chat-HD achieves performance matching the TSM baseline on the MoLa dataset demonstrating that a general-purpose VLM can match a task-specific model. However, when evaluated on the generated dataset, TSM's performance drops significantly, with recall decreasing to 4.2%. This may be due to its dependence on motion-based features, the relatively short duration of violent actions in the generated dataset, and its limited ability to generalize effectively. In contrast, IV2-Chat-HD maintains strong performance, achieving a recall of 99.1% on the unseen dataset. This suggests that VLMs learn more generalizable representations, possibly because language alignment encourages the model to capture the semantic meaning of violence rather than relying solely on motion patterns allowing it to detect even subtle forms of violence. Overall, these results indicate that, in our experimental setting, the VLM IV2-Chat-

HD demonstrates better generalization compared to the CNN-based TSM.

## VI. Conclusion

In this study, we investigated the transition from traditional supervised architectures to multimodal Video-Language Models (VLMs) for the task of in-car violence detection in autonomous vehicles. Our experiments demonstrated that while traditional models like TSM achieve high accuracy on same-distribution data, they suffer from severe generalization failures when exposed to new distributions, as evidenced by a drop to 4.2% recall on the synthetic dataset. In contrast, the fine-tuned InternVideo2-Chat model not only matched the performance of the TSM baseline on the MoLa dataset but also showed stronger generalization.

Furthermore, our work highlights the transformative potential of generative AI in addressing the scarcity of sensitive datasets. By utilizing the Wan2.2 diffusion model, we showed that synthetic data can effectively supplement or reduce reliance on real-world video data for fine-tuning.

Overall, these findings indicate that combining VLMs with synthetic training data is a promising direction for in-car safety monitoring. Although further validation on larger and more diverse real-world datasets is required, the proposed approach provides a practical foundation for developing reliable violence detection systems under limited data availability. Future research will focus on scaling the synthetic dataset using more advanced closed-source generative models, as well as optimizing VLMs to be lightweight enough for deployment on edge devices deployed in vehicles, rather than relying on cloud-based systems.

## References


[1] J. Zhao, Y. Wu, R. Deng, S. Xu, J. Gao, and A. Burke, "A survey of autonomous driving from a deep learning perspective," *ACM Computing Surveys*, vol. 57, no. 10, pp. 1–60, 2025.

[2] D. Parekh, N. Poddar, A. Rajpurkar, M. Chahal, N. Kumar, G. P. Joshi, and W. Cho, "A review on autonomous vehicles: Progress, methods and challenges," *Electronics*, vol. 11, no. 14, p. 2162, 2022.

[3] S. Waykole, N. Shiwakoti, and P. Stasinopoulos, "Review on lane detection and tracking algorithms of advanced driver assistance system," *Sustainability*, vol. 13, no. 20, 2021.

[4] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4724–4733.

[5] R. Goyal, S. E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, and R. Memisevic, "The "something something" video database for learning and evaluating visual common sense," 10 2017, pp. 5843–5851.

[6] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 7082–7092.

[7] Y. Wang, K. Li, X. Li, J. Yu, Y. He, C. Wang, G. Chen, B. Pei, Z. Yan, R. Zheng, J. Xu, Z. Wang, Y. Shi, T. Jiang, S. Li, H. Zhang, Y. Huang, Y. Qiao, Y. Wang, and L. Wang, "Internvideo2: Scaling foundation models for multimodal video understanding," 2024.

[8] L. H. Backar, M. A. Khalifa, and M. A.-M. Salem, "In-vehicle monitoring for passengers' safety," in *2022 IEEE 12th International Conference on Consumer Electronics (ICCE-Berlin)*, 2022, pp. 1–6.

[9] A. M. Abuomar, Y. A. Ahmed, and M. A.-M. Salem, "Safety on wheels: Computer vision for driver and passengers monitoring," in *2023 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, 2023, pp. 29–34.

[10] Y. Albadawi, M. Takruri, and M. Awad, "A review of recent developments in driver drowsiness detection systems," *Sensors*, vol. 22, no. 5, 2022.

[11] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6479–6488.

[12] E. B. Nievas, O. D. Suarez, G. B. García, and R. Sukthankar, "Violence detection in video using computer vision techniques," in *Computer Analysis of Images and Patterns: 14th International Conference*. Springer, 2011, pp. 332–339.

[13] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," in *IEEE computer society conference on computer vision and pattern recognition workshops*, 2012, pp. 1–6.

[14] N. R. Rodrigues, N. M. da Costa, R. Novais, J. Fonseca, P. Cardoso, and J. Borges, "Ai based monitoring violent action detection data for in-vehicle scenarios," *Data in brief*, vol. 45, p. 108564, 2022.

[15] N. R. Rodrigues, N. M. da Costa, C. Melo, A. Abbasi, J. C. Fonseca, P. Cardoso, and J. Borges, "Fusion object detection and action recognition to predict violent action," *Sensors*, vol. 23, no. 12, p. 5610, 2023.

[16] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[17] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6201–6210.

[18] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 20–36.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[20] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[21] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.

[22] Y. Wang, K. Li, Y. Li, Y. He, B. Huang, Z. Zhao, H. Zhang, J. Xu, Y. Liu, Z. Wang, S. Xing, G. Chen, J. Pan, J. Yu, Y. Wang, L. Wang, and Y. Qiao, "Internvideo: General video foundation models via generative and discriminative learning," 2022.

[23] T. Wan, A. Wang, B. Ai, B. Wen, C. Mao, C.-W. Xie, D. Chen, F. Yu, H. Zhao, J. Yang *et al.*, "Wan: Open and advanced large-scale video generative models," *arXiv preprint arXiv:2503.20314*, 2025.

[24] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4340–4349.

[25] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid, "Learning from synthetic humans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[26] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2758–2766.

[27] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*. PMLR, 2017, pp. 1–16.

[28] A. Ullah, M. Numan, M. N. A. Khalid, and A. Majid, "Words shaping worlds: A comprehensive exploration of text-driven image and video generation with generative adversarial networks," *Neurocomputing*, vol. 632, p. 129767, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231225004394

[29] W. Li, D. Luo, D. Yang, Z. Li, W. Wang, and Y. Zhou, "The role of video generation in enhancing data-limited action understanding," *arXiv preprint arXiv:2505.19495*, 2025.