Appendix

In this appendix, we provide the following material:

- More detailed implementation details (see Appendix A).
- Additional comparison between models pre-trained with different Masked Video Modeling (MVM) objectives (see Appendix B).
- Detailed evaluation and comparison with existing methods (see Appendix C).
- Annotation refinement (see Appendix D).

A. Implementation details

Encoder-only architecture. With our simple encoder-only design, we pass a video input $X_t \in \mathbb{R}^{T\times H\times W\times 3}$ at time t to a Video Foundation Model (ViFM) encoder, obtaining a spatio-temporal feature vector $F_t \in \mathbb{R}^E$, and then transform it into classification logits $L_t \in \mathbb{R}^C$ using one linear layer.

The way F_t is obtained depends on the specific encoder design. For Video ViTs, we retain the original architectures used for general video classification. First, the input is divided into tubelet tokens (spatio-temporal patches), which are then passed through a series of Transformer blocks. Then, the resulting tokens are aggregated into one spatio-temporal embedding. All Video ViT models we used, except for InternVideo2 [52], do not include a dedicated classification token and apply mean pooling over the resulting token embeddings. InternVideo2 [52] includes an additional token, but instead of treating it separately, the model aggregates all the resulting tokens using an attention pooling layer, which computes a weighted sum of the embeddings of the token, allowing the model to adaptively focus on the most informative spatio-temporal regions. Finally, as in the original designs, the aggregated feature vector is further processed with a final Layer Normalization [2] operation to stabilize and standardize the representation. For the R(2+1)D model [45], which follows a convolutional design, no explicit token aggregation step is required. The model processes the spatio-temporal input through 3D convolutional layers and directly produces a single spatio-temporal embedding. We adopt the implementation from CycleCrash [8], where the final spatio-temporal embedding is passed through a batch normalization layer and ReLU activation to obtain the final spatio-temporal feature vector F_t.

Domain-adaptive pre-training. We base our training on the VideoMAE [44] pretraining recipe with AdamW[29] optimizer and MSE loss on masked tokens. We downscale the training and set the batch size to 800 with 1M samples per epoch, which corresponds to 1250 iterations per epoch. We set the cosine learning rate schedule for 20 epochs with 1 epoch of linear warmup, but stop after 12 epochs by default. We mask out 75% of the frames.

When mixing normal and abnormal driving data, each batch consists of 480 samples (60%) from the normal driving dataset BDD100K [55], and 320 samples (40%) from the abnormal driving dataset CAP-DATA [11]. Note that CAP-DATA includes both normal and abnormal driving examples.

Fine-tuning. All Video ViT models are fine-tuned under the same configuration using a batch size of 56. We closely follow the VideoMAE [44] fine-tuning recipe for the small HMDB51[21] dataset with AdamW[29] optimizer, cosine learning rate schedule, and the cross-entropy loss. We set the layer decay rate to 0.6, take 50K training examples per epoch (randomly chosen each new epoch), and fine-tune for 50 epochs with 5 epochs of linear warm-up. For the Base and Large variants, we also set the learning rate to 5e-4.

For MOVAD [37], we adopt the official publicly available implementation. For VidNeXt [8], its ablations, and the R(2+1)D model [45], we reimplement training within our pipeline to ensure consistency across datasets. Where applicable, we align fine-tuning hyperparameters with the original recipes, while training these models for the same number of iterations as Video ViT models.

Evaluation. We report AUC_{ROC} using checkpoints with the highest AUC_{ROC} on the validation set. For AUC_{MCC} and MCC@0.5, we use checkpoints with the highest AUC_{MCC} on the validation set.

B. Effectiveness of different MVM objectives for TAD

We demonstrated in Sec. 4.3 and Tab. 3 that among fully-supervised (FSL), weakly-supervised (WSL), and self-supervised (SSL) pre-training, the latter is the most suitable for TAD. Self-supervised pre-training methods for video are predominantly based on the MVM approach, which reconstructs masked video patches. However, specific designs of the MVM pre-training approach vary in how patches are masked and what is reconstructed.

To better understand the effect of specific MVM objectives and masking strategies, we select several ViFMs that differ only in these parameters and fine-tune them on DoTA [54]. All models were pre-trained on Kinetics-400 [17] for 1600 epochs with the spatio-temporal patch size of 2x16x16 and the same input size of 16x224x224.

VideoMAE [44] employs random tube masking, where contiguous spatio-temporal volumes are randomly masked to encourage learning from structured visual dynamics. Being a Masked Autoencoder (MAE) method, it reconstructs RGB pixels of masked patches. MME [41] introduces motion-aware masked autoencoding by reconstructing dense motion trajectories instead of raw pixels. It randomly masks tubes and trains the model to predict motion features extracted from frame differences, promoting a stronger understanding of dynamic content in videos. SIGMA [38] aims to learn high-level semantics with a pro-

Model	AUC_{ROC}	AUC_{MCC}	MCC@0.5	Masking strategy	Reconstruction objective
VideoMAE [44]	86.3	54.8	57.8	Random tube	Pixel
MME [41]	86.3	**55.2**	57.8	Random tube	Motion trajectory
SIGMA [38]	86.4	54.8	57.9	Random tube	Cluster assignments
MGMAE [15]	**86.6**	55.0	**58.2**	Optical flow guided token	Pixel

Table 5. **Effect of MVM pre-training objectives.** Despite variations in masking strategy and reconstruction objectives, all MVM-based models show strong performance, indicating that MVM is broadly effective for TAD. Models leveraging motion-aware objectives or masking (MME [41], MGMAE [15]) slightly outperform those relying on raw pixels or semantics (VideoMAE [44], SIGMA [38]), suggesting that motion modeling benefits TAD, even when pre-trained on general datasets. Using Video ViT-Base models initialized with weights pre-trained on Kinetics-400 [17] and fine-tuned on DoTA [54].

jection network, so instead of raw pixels, it reconstructs semantic cluster assignments derived via optimal transport over spatio-temporal features. On the opposite, MGMAE [15] keeps raw pixels as its reconstruction objective and focuses on advancing the masking strategy. It introduces a motion-guided masking mechanism that leverages optical flow to prioritize masking regions with higher motion, encouraging the model to focus on dynamic and informative parts of the video.

We compare the performance of these models on DoTA in Tab. 5 and see that, despite using different objectives and masking strategies, all MVM-based models achieve similarly strong performance on TAD, confirming that MVM pre-training is robust and broadly effective for this task. Notably, MME slightly outperforms others in AUC_{MCC} by predicting motion trajectories rather than pixels, suggesting that incorporating motion dynamics into the reconstruction objective may help the model better capture temporal cues relevant for anomaly detection. MGMAE achieves the highest AUC_{ROC} and MCC@0.5, indicating that motion-guided masking can help the model focus on dynamic and informative regions. In contrast, SIGMA, which reconstructs high-level semantic clusters, performs on par with VideoMAE, providing no clear evidence that high-level semantics improve TAD performance.

These results indicate that motion-oriented objectives and masking strategies provide consistent benefits for TAD, even when pre-training is performed on general video datasets. In contrast, high-level semantic reconstruction shows only marginal gains, suggesting that focusing on such features may not directly benefit tasks requiring fine-grained temporal reasoning like TAD, unless better aligned with the demands of the task.

C. Detailed evaluation and comparison with existing methods

We present additional experiments that compare our encoder-only ViFM-based models with top-performing specialized TAD methods and analyze key aspects of their performance more closely.

Short	Anomaly Categories
ST	Collision with another vehicle which starts, stops, or is stationary
AH	Collision with another vehicle moving ahead or waiting
LA	Collision with another vehicle moving laterally in the same direction
OC	Collision with another oncoming vehicle
TC	Collision with another vehicle which turns into or crosses a road
VP	Collision between vehicle and pedestrian
VO	Collision with an obstacle in the roadway
OO	Out-of-control and leaving the roadway to the left or right
UK	Unknown

Table 6. **Traffic anomaly categories in the DoTA dataset.** Each category is represented by scenarios with and without ego-car participation.

First, we compare our models with existing specialized TAD methods across anomaly categories and groups of the DoTA [54] dataset. DoTA comprises scenarios where the ego-vehicle, from which the video is captured, is either involved in the anomaly or observes other road agents involved in it. We show the list of anomaly categories in DoTA in Tab. 6.

We show the results across categories in Tab. 7. We can see that for most of the categories, and especially for the ego-involved group, our DAPT-VideoMAE models outperform or are on par with specialized methods. ISCR-TAD [25] and PromptTAD [35] frequently rank among the top three methods across several categories. That suggests that incorporating object information and various related representations, such as depth or high-frequency features, is beneficial, especially in highly untypical accident scenarios (categories UK and UK*).

Next, in Tab. 8, we assess the generalization performance of the models by evaluating those trained on DoTA [54] directly on the DADA-2000 [10] dataset. VideoMAE-based models already outperform specialized methods by a large margin, and applying DAPT further improves performance.

Ego-vehicle involved video clips										
Method	ST	AH	LA	OC	TC	VP	VO	OO	UK	AVG
STFE [57]	75.2	84.5	72.1	77.3	72.8	71.9	–	–	–	75.6
MOVAD [37]	86.6	86.3	84.9	83.7	85.5	81.6	77.4	87.9	73.8	83.1
TTHF [24]	86.7	90.5	89.7	87.0	89.5	77.1	87.6	90.1	70.9	85.5
PromptTAD [35]	84.2	90.2	88.4	85.6	89.1	**83.6**	86.8	88.7	74.6	–
ISCRTAD [25]	81.7	89.2	89.9	87.4	90.9	83.1	**90.1**	88.9	**78.9**	86.7
DAPT-VideoMAE-S	**87.3**	**91.1**	**90.2**	**87.6**	**91.0**	**83.2**	85.7	**91.2**	75.8	**89.9**
DAPT-VideoMAE-B	**87.5**	**92.0**	**90.7**	**89.6**	**91.8**	81.6	**88.4**	**91.1**	**76.1**	**90.7**
DAPT-VideoMAE-L	**87.4**	**90.6**	**91.4**	**88.7**	**91.8**	**85.9**	**89.8**	**91.6**	**76.0**	**90.7**
Ego-vehicle NOT involved video clips										
	ST*	AH*	LA*	OC*	TC*	VP*	VO*	OO*	UK*	AVG*
STFE [57]	**80.6**	65.6	69.9	76.5	74.2	–	75.6	70.5	–	73.2
MOVAD [37]	72.2	74.0	74.8	80.2	79.6	76.8	82.2	78.3	72.9	76.8
TTHF [24]	74.9	76.0	76.4	79.8	81.5	79.2	79.0	77.5	68.9	77.0
PromptTAD [35]	73.8	**78.7**	**81.8**	82.8	**85.1**	84.6	83.1	82.4	**79.1**	–
ISCRTAD [25]	**84.6**	**78.7**	77.3	85.8	82.5	**86.8**	**85.4**	**84.5**	73.5	**82.1**
DAPT-VideoMAE-S	78.9	77.8	78.1	**86.6**	83.9	82.9	78.6	81.8	**75.7**	81.6
DAPT-VideoMAE-B	**80.2**	**80.0**	**83.6**	**86.6**	**85.8**	**86.7**	**84.6**	**84.4**	73.0	**84.1**
DAPT-VideoMAE-L	79.0	**81.4**	**86.1**	**88.2**	**86.3**	**85.8**	**85.7**	**85.6**	**75.6**	**85.2**

Table 7. **Comparison with specialized TAD methods across categories.** Our simple encoder-only models outperform or match top specialized methods across majority of categories. Models using explicit object or scene cues and extra representations (*e.g.*, ISCRTAD [25], PromptTAD [35]) show advantage in rare or ambiguous scenarios, such as UK and UK*. Reporting AUC_{ROC} (%) of individual accident categories on the DoTA [54] dataset.

Method	**Ego**	**Non-Ego**	**Both**
TTHF [24]	80.9	64.0	71.7
PromptTAD [35]	79.7	70.4	74.6
ISCRTAD [25]	82.7	66.9	74.2
VideoMAE-S_{HALF}	89.6	74.8	81.6
VideoMAE-S	90.2	78.3	83.7
VideoMAE-B	91.2	79.6	84.9
VideoMAE-L	92.0	81.0	86.2
DAPT-VideoMAE-S_{HALF}	90.4+0.8	77.5+2.7	83.4+1.8
DAPT-VideoMAE-S	91.2+1.0	78.7+0.4	84.3+0.6
DAPT-VideoMAE-B	92.4+1.2	80.3+0.7	85.8+0.9
DAPT-VideoMAE-L	91.9 -0.1	82.3+1.3	86.6+0.4

Table 8. **Generalization performance.** VideoMAE-based [44] models generalize significantly better than specialized methods, and even a lightweight variant trained on only half of DoTA [54] outperforms prior work, underscoring the robustness of foundation model pretraining. AUC_{ROC} (%) of different methods trained on DoTA [54] and evaluated on DADA-2000 [10].

VideoMAE-S_{HALF}, a small variant of a Video ViT fine-tuned on only half of the DoTA dataset, already outperforms specialized methods significantly, highlighting the strong generalization capabilities commonly attributed to foundation models.

D. Annotation refinement

While DoTA [54] provides large-scale and comprehensive annotations for traffic anomaly detection, we noticed minor inconsistencies and structural issues that could affect training and evaluation. To identify these cases, we fine-tuned a VideoMAE-Small [44] model on the official training set

and calculated the error rate for each video clip in the set. After that, we flagged the clips with the highest error rate. This model-guided filtering exposed potentially mislabeled or ambiguous samples. Manual review revealed issues, such as substantially imprecise anomaly timing and distinct clips merged into a single video file, which we manually corrected. We repeated the same procedure (including fine-tuning) on the validation set. Although only about 1% of the clips were refined, we release the corrected annotations for reproducibility. We use the refined DoTA dataset for comparisons between Video ViT models, and the original dataset when comparing to existing specialized methods.

We repeated the same process for DADA-2000 [10], but a brief manual review did not reveal any significant inconsistencies.