A Lifesaving Foundation: Leveraging Foundation Models for Recognizing and Anticipating Medical Procedure Actions

Adrian L. Lindloff[1][0009−0009−3434−0058], Simone Frintrop[1][0000−0002−9475−3593], Christian Wilms[1*][0009−0003−2490−7029], and Matthias Wilms[2*][0000−0001−8845−360X]

[1] Computer Vision Group, University of Hamburg, Hamburg, Germany
adrian.lindloff@studium.uni-hamburg.de, simone.frintrop@uni-hamburg.de, christian.wilms@uni-hamburg.de
[2] Department of Radiology, University of Michigan, Ann Arbor, USA
wilms@umich.edu
* denotes shared senior authorship

Abstract. First responders play a pivotal role in medical emergency situations, where timely and accurate decision-making can be lifesaving. To provide support in such scenarios, computer vision-based systems with the ability to recognize the currently ongoing procedure and anticipate the next steps are of significant interest, as they could assist in guiding first responders through critical interventions. However, annotated data to train such systems is scarce. Therefore, we study the applicability of recent foundation models on the tasks of recognizing and anticipating medical procedure actions from video data as part of The Trauma THOMPSON Challenge 2025. Specifically, we adapt image-based DINOv2, video-based VideoMAE, and multi-modal ActionCLIP to the problem by adjusting prompt handling, hyperparameters, and loss weighting. In our evaluation, we show that despite the limited amount of training data in the challenge dataset, the models are able to produce decent results in the recognition of medical procedure actions. Moreover, they demonstrate strong potential, indicated by improved performance under less strict evaluation criteria[1].

Keywords: Action Recognition · Action Anticipation · Life-Saving Interventions (LSI) · Foundation Models · Video Classification.

1 Introduction

In medical emergency situations, first responders play a critical role and often dictate the outcome regarding not just the immediate, but also the long-term future of a patient [10]. First responders need to quickly identify the precise characteristics of such situations and perform the appropriate medical treatment under immense pressure. Support from an experienced assistant can substantially

[1] https://github.com/Loffl/A-Lifesaving-Foundation-The-Trauma-THOMPSON-Challenge-2025-solution

improve the quality; however, such support is rarely available due to the nature of these situations. Therefore, computer vision-based autonomous remote instruction systems are envisioned to mitigate the lack of human assistance.

With this goal in mind, the Trauma THOMPSON Challenge 2025 [22] was proposed to facilitate the development of such systems for lifesaving interventions. The challenge introduces a first-person video dataset of unscripted emergency interventions to foster the development of algorithms that can recognize and anticipate medical procedures, detect tools, and track hand movements during an intervention, assess the realism of a simulated trauma procedure, and perform visual clinical question answering. Here, we focus on the recognition and prediction of medical procedure actions, both tasks that are closely related to general action recognition in computer vision. While action recognition refers to assigning an action to the currently analyzed video part, action prediction involves forecasting the action of the subsequent video part.

Over the years, many approaches have been proposed to address the video action recognition task. First, before public video datasets of sufficient size existed, methods such as cuboids [6] relied on traditional feature engineering to design hand-crafted spatio-temporal representations to detect actions. Using deep learning, architectures evolved to be more automated and, among other techniques, (1) extract per frame features through 2D CNNs [18], also with an additional LSTM on top [7], (2) apply 3D CNNs directly to videos [4,17], (3) extract spatial and temporal features individually with two-stream networks [9], or more recently, (4) apply vision transformers (ViTs) [8] to the video classification problem [1]. In recent years, foundation models have substantially improved the performance across various computer vision domains, including video-related tasks such as video classification [2,20]. Due to their extensive training on huge amounts of data, foundation models are able to generate rich general-purpose features, allowing them to generalize well to a wide range of tasks without extensive training data. This makes foundation models highly suitable for tasks where data is expensive or difficult to acquire, like in medical applications [13].

In this paper, we adapt three well-known foundation models: DINOv2 [14], VideoMAE [16], and ActionCLIP [19], to the task of medical procedure action recognition and anticipation in the context of the Trauma THOMPSON (T3) Challenge 2025. Foundation models with their strong few-shot capabilities [2] may be particularly well-suited for the setting of the 2025 T3 Challenge, which features 124 classes in a long-tailed distribution across 2477 training samples. The three selected foundation models cover a variety of input modalities, including images, video, and text (see Fig. 2), which allows us to study the effect of these modalities. To adapt the foundation models for the dual-task formulation, we incorporate video action classifiers for the recognition and anticipation tasks when necessary. In our evaluation of the 2025 T3 Challenge data, the three models with tailored prompt handling, hyperparameters, and loss weighting demonstrate promising results.

2 Dataset

The T3 Challenge 2025 dataset [22], which we use in this work, is an egocentric video dataset specifically built for medical procedure action recognition and anticipation. For this task, corresponding to task 1 in the challenge, it comprises 141 videos of regular life-saving intervention procedures in its training set and another 79 videos in the test set. The test set is further split into 36 unseen regular procedure videos and 43 just-in-time procedure videos. Each video presents a medical procedure and consists of several shorter video clips, representing individual actions. In total, the training dataset consists of 2477 video clips. On average, a single video clip consists of around 152 frames, and the original resolution of the video clips is 1920×1080. Following the official challenge instructions, we resized the video clips to 480×270 pixels.

Each video clip is annotated with one action label representing the currently performed action. Each label consists of one verb and one noun, which individually can reappear in other action labels. Each clip annotation also contains the action label of the subsequent video clip. In total, there are 124 action classes, composed of 42 unique nouns and 42 distinct verb classes. The class distribution represents the real-world distribution of the procedures, similar to how the procedures themselves were chosen to reflect relevant medical interventions. This leads to a long-tailed class distribution, as illustrated in Fig. 1. As a consequence, a few common action classes dominate a significant number of minority classes. For reference, the most frequent action 'insert finger' appears in 85 video clips, while a multitude of actions only appear a single time.

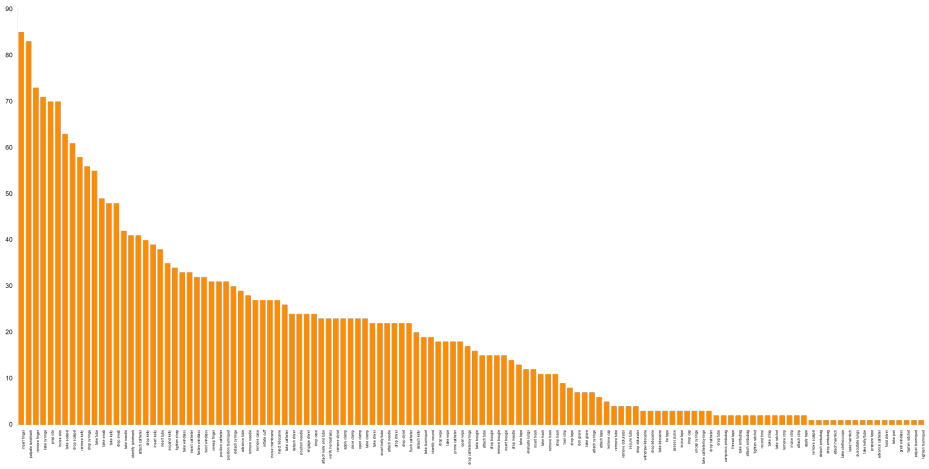
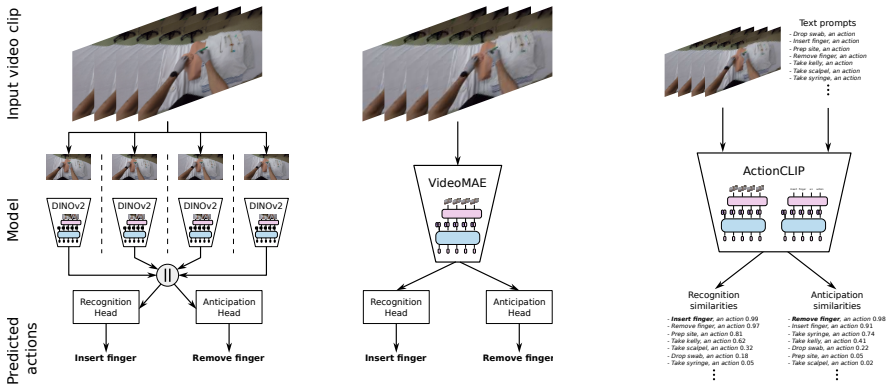


Fig. 1: Distribution of action classes (combinations of nouns and verbs) in the training set of the T3 Challenge 2025 dataset, with actions on the x-axis and frequency on the y-axis.

For our work, we split the original training set into our training set (80%) and our validation set (20%). Hence, 1968 video clips remain for the training. Due to the considerable class imbalance, 16 of the 124 action classes occur only once and can not occur in both sets. To enable the model to learn from all classes, these samples are added to our training set. Consequently, our validation set only includes 108 action classes.



(a) DINOv2 (b) VideoMAE (c) ActionCLIP

Fig. 2: General processing of videos for action recognition and anticipation with the three foundation models. Image-based DINOv2 (a) processes individual frames and combines the results in the final stage. Video-based VideoMAE (b) directly extracts features from the entire video clip with subsequent classification. Multi-modal ActionCLIP (c) extracts features from the entire video and the label via text prompt. Subsequently, classification is performed as a similarity calculation between a video and all possible text prompt representations.

3 Method

This section introduces our adapted versions of the three foundation models, which we employ for action recognition and anticipation (task 1) of the T3 Challenge 2025. For each model, we discuss the architecture and the implementation details. The presented configurations are based on the best-performing results obtained through extensive evaluation on the 2023 T3 Challenge dataset.

3.1 DINOv2

DINOv2 [14] is a discriminative self-supervised learning approach, which produces general-purpose visual features for images. Besides other computer vision tasks, DINOv2 was successfully applied to video action recognition tasks outside the medical domain [14]. Among self-supervised approaches at the time

of its publication, DINOv2 achieved state-of-the-art accuracy performance even though the model was trained on images only. Therefore, we adapt DINOv2 to the T3 Challenge 2025.

Architecture The DINOv2 [14] learning approach combines the ideas of DINO [3] and iBOT [21]. DINO [3] works at the image-level, after obtaining features from the ViT backbone. It utilizes a student-teacher architecture, where each receives a different crop of the same image as input to generate a ViT class token. Each token is forwarded to the respective head for loss calculation.

The iBOT method [21] is employed at the patch-level and thus alters the input behavior of the ViT model. The input image patches are randomly masked and distributed to the same student and teacher model. The student operates on the masked patches, whereas the teacher employs the entire tokenized image without any masks, enabling the teacher to process information unavailable to the student network. The output features are passed to the corresponding student and teacher iBOT MLP head to produce the loss.

At first, both the DINO and iBOT methods are strictly separated from each other as they compute their own losses. Their knowledge is only brought together when both losses of a given batch are combined with a regularization term to train the student model. After the pre-training has been completed, only the student network is used to generate image representations.

To apply DINOv2 to video action recognition, we obtain the output embeddings for each frame as depicted in Fig. 2a. The classification (`CLS`) tokens, as well as the remaining embedding sequences of all frame representations for a sample, are averaged and then concatenated to imitate video representations. These are forwarded to two linear classification heads for recognition and anticipation, respectively (see bottom part of Fig. 2a). The Cross-Entropy Loss is applied to the resulting logits to compute the action losses.

Concerning anticipation, only valid samples are considered for the loss computation. A sample is not valid when there is no next action present, in case it is the last clip of a video. Both the recognition and the anticipation losses are combined to form a final action loss, which, in the simplest case, is obtained by weighting them equally. In our experiments, we investigate the impact of different loss weighting methods on the model's performance.

Implementation Details To apply DINOv2, we first uniformly sample 16 frames for each video clip. Next, each frame is randomly cropped and resized to 224×224. A higher resolution is not possible when using the pre-trained model weights. The frames are horizontally flipped with a 50% probability and randomly rotated by up to two degrees, before being scaled by a factor between 0.8 and 1.2. The validation and test data are only uniformly sampled and resized.

The following parameters were utilized for our DINOv2 experiments on the 2025 T3 Challenge dataset. The batch size is set to 20, and training lasts for 50 epochs. Gradient clipping is enabled and set to 1.0. For the optimizer, AdamW is employed with a β_1 of 0.9, a β_2 of 0.999, and $\epsilon = 10^{-8}$. Weight decay is set to

0.1, and the learning rate is initialized at 5×10^{-5}. It is controlled by a cosine annealing scheduler with a linear warm-up phase over the first 10% of training steps. Gradient accumulation is enabled and fixed at 2, meaning that gradients are accumulated for two forward passes before the weights are updated. We use a large-sized ViT backbone with four additional register tokens. The register tokens [5] have proven to sufficiently remove existing artifacts in the Vision Transformer attention maps.

3.2 VideoMAE

VideoMAE [16] by Tong et al. is a transformer-based foundation model, explicitly designed to process video data (see Fig. 2b) as it extends the concept of masked autoencoders for self-supervised pre-training to videos. Similar to the DINOv2 approach for images, VideoMAE extracts semantically rich general-purpose video representations. Since it is designed to extract features from videos, action recognition constitutes an intuitive application, which Tong et al. briefly explore, reporting state-of-the-art accuracies on relevant benchmarks.

Architecture Inspired by ImageMAE [11], VideoMAE performs a masking and reconstruction task using an asymmetric encoder-decoder architecture in a self-supervised setup. The encoder is a ViT with joint space-time attention. Before a video is processed by the encoder, uniform temporal downsampling is applied. Subsequently, the video is embedded into 3D tokens using joint space-time cube embedding, which are then mapped to the transformer embedding dimension. Each cube has the size $2 \times 16 \times 16$, meaning it contains spatial patches of two frames. Next, tube masking is applied. Due to the high redundancy in video data, the masking ratio is set between 90% and 95%. The non-masked tokens are used as the input for the encoder. Given the output representations of the visible tokens from the encoder and the remaining learnable masked tokens, a lightweight ViT-based decoder reconstructs the missing tokens and thus recreates the entire video. This enables VideoMAE to learn meaningful video representations. The generated video representations are processed by two linear classifiers as visualized in the bottom part of Fig. 2b, one for obtaining recognition logits and another for anticipation probabilities when a sample is valid.

Implementation Details To adapt VideoMAE to the T3 Challenge 2025, we uniformly downsample each video to 16 input frames. We then randomly downsize each frame while maintaining the original aspect ratio and randomly extract crops of size 224×224. Finally, random left-right flipping is applied. Remaining transformations correspond to the DINOv2 data preprocessing. For the validation and test data, only the initial temporal downsampling is applied. Hence, we do not apply test-time augmentation.

The VideoMAE configurations of our experiments adopt the following training parameters. We set the batch size to 2 and train for 50 epochs. There is no

weight decay, and the learning rate is set to 5×10^{-5}. The ViT backbone is huge-sized, and no register tokens are utilized. All remaining configuration settings mirror the training method of DINOv2, including the Cross-Entropy Loss.

3.3 ActionCLIP

The CLIP-based [15] model ActionCLIP [19] is a foundation model proposed by Mengmeng Wang et al. It was specifically designed for video action recognition and could thus translate particularly well to the target data. It has outperformed other ViT-based model architectures on the Kinetics-400 [12] action recognition benchmark, setting new state-of-the-art accuracy scores.

Architecture In contrast to DINOv2 and VideoMAE, which exclusively rely on visual input, ActionCLIP [19] utilizes both visual and text supervision in a multi-modal framework tailored for action recognition. It frames action recognition as a video-text similarity matching problem. To achieve this, ActionCLIP creates distinct representations for the visual and text input using two encoders as depicted in Fig. 2c.

The text encoder is a 12-layer Transformer based on the CLIP architecture [15]. As input prompts for the text encoder, we utilize 16 possible templates for each of the 124 action classes in the 2025 T3 Challenge. Given a label, a prompt is created by replacing a blank field within a template, for example, "*Label*, an action" for the recognition task. When anticipating actions, we apply slightly modified templates such as "*Label*, an upcoming action". During the training, a single text prompt is randomly selected from the 16 available candidates for each video clip and for action recognition and anticipation labels, respectively, which is then fed to the text encoder. For inference, all 16×124 prompts are used for both subtasks, with the most similar prompt indicating the predicted action label, as visualized in the bottom part of Fig. 2c, which visualizes this process with seven example prompts.

The vision encoder in ActionCLIP is a standard ViT-Base with a patch size of 16×16. We use ActionCLIP's post-network prompt to capture temporal relations between different frames. This is accomplished by adding a second, separate vision encoder, which receives the concatenated frame-level representations of the initial pre-trained CLIP vision encoder. The second encoder is a temporal transformer with extra learnable temporal positional embeddings and a mean-pooled output sequence, yielding a feature representation of the entire video.

Given the text and vision representations, ActionCLIP calculates the similarities between them. The idea is to align matching representations, a text prompt that describes the action of the respective video, by pulling the pairwise text and video representations close to each other. This is achieved by computing symmetric cosine similarity scores across all conceivable modality pairs of feature embeddings. The video-to-text similarity scores quantify how similar a given video embedding is to all other text embeddings in the batch, while text-to-video scores capture the reverse relationship.

Implementation Details To handle the T3 Challenge 2025 video data with ActionCLIP, we split every video clip into 16 segments, from which two consecutive frames are sampled each. We largely follow Wang et al. [19] to pre-process the extracted frames. First, two random image augmentations are applied to each frame, given a pool of 14 possible operations, including geometric, color, and intensity transformations [**?**]. Subsequently, the frames are cropped randomly at different scales and resized to 224×224 pixels. Finally, the frames are horizontally flipped with a probability of 50%, a random color jitter is applied, and 20% of the frames are randomly converted to grayscale. All frames of a video are stacked and provided as input to ActionCLIP. During validation and testing, the frames are only resized and center-cropped. Hence, we do not apply test-time augmentation.

To train ActionCLIP, the Kullback-Leibler (KL) divergence is utilized as a loss function. Given a label y and a paired video clip x, it is calculated once for the video-to-text similarity scores $p^{x2\boldsymbol{y}}(\boldsymbol{x})$ and once for the text-to-video similarity scores $p^{y2x}(\boldsymbol{y})$ of a batch, to create a symmetric contrastive learning objective that is defined as:

$$\mathcal{L} = \frac{1}{2}\mathbb{E}_{(x,y)\sim\mathcal{T}}\left[\mathrm{KL}\left(\boldsymbol{p}^{x2y}(\boldsymbol{x}), \boldsymbol{q}^{x2y}(\boldsymbol{x})\right) + \mathrm{KL}\left(\boldsymbol{p}^{y2x}(\boldsymbol{y}), \boldsymbol{q}^{y2x}(\boldsymbol{y})\right)\right] \tag{1}$$

where $\mathcal{T}$ denotes the entire training set and $q^{x2y}(x), q^{y2x}(y)$ correspond to the ground-truth similarity scores [19]. Since the recognition and the anticipation tasks employ different text embeddings, we calculate four similarity scores and two losses. We evaluate several strategies to combine the recognition and anticipation losses in Sec. 4.1.

Our ActionCLIP experiments utilize a batch size of 12, the training lasts for 75 epochs, and gradients are clipped when exceeding a value of 20. AdamW is the optimizer, with a β_1 of 0.9, a β_2 of 0.98, and $\epsilon = 10^{-8}$. The weight decay rate is set to 0.2 and the base learning rate to 5×10^{-6}. The post-network prompt is trained with a learning rate set to ten times the base rate. A standard cosine annealing scheduler with a five-step linear warm-up is employed.

4 Results

In this section, we evaluate the three foundation models on the T3 Challenge 2025 dataset to assess their suitability for the action recognition and anticipation tasks. All models and variations are trained on our training set as outlined in Sec. 2 and evaluated on our validation set or the official test set. Top-1 and top-5 accuracies serve as the evaluation metrics for both tasks individually. We also report the average accuracy across both tasks to evaluate the overall performance. Note that on the official test set, only top-1 accuracies are available.

Loss Weighting Since task 1 of the 2025 T3 Challenge is formulated as a dual-task problem, consisting of action recognition and anticipation, we compare three strategies for weighting the individual task losses, yielding the overall loss. First, we combine the two losses for action recognition and anticipation by scaling them with 0.5 before adding them. We coin this strategy *Equal*. Second, we progressively adapt the loss weights throughout the training process. Specifically, we start with a higher weight on the recognition loss (0.6) compared to the anticipation loss (0.4) and linearly adjust the weights between epochs such that in the final epoch, we again reach 0.5 for both losses. This strategy is called *Linear* and follows the premise that recognition might be easier to learn initially and, therefore, stronger supervision could result in better shared representations. As our third strategy, we follow Wang et al. [19] and apply the *Equal* strategy, but introduce an additional loss term scaled by a factor that linearly decreases from 0.15 to 0.05 throughout training. The additional loss is a symmetric bidirectional KL Divergence loss, which encourages the model to predict similar distributions for recognition and anticipation, as both actions are often closely related in their temporal context. Hence, we name the strategy *KLD*.

4.1 Results on the Validation Set

Table 1: Results of the three foundation models on our validation set for the T3 Challenge 2025, employing each of our three loss weighting strategies. Top-1 and top-5 accuracies are reported separately for recognition, anticipation, and the combination. Bold font highlights the best overall results, whereas slanted font indicates the best results for each foundation model.

	Loss	Recognition		Anticipation		Total	
Model	Weighting	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
DINOv2	Equal	42.83%	80.55%	25.34%	76.23%	34.09%	78.39%
VideoMAE	Equal	46.17%	74.66%	13.95%	64.24%	30.06%	69.45%
ActionCLIP	Equal	*40.48%*	*82.74%*	***35.15%***	*82.01%*	***37.82%***	*82.38%*
DINOv2	Linear	41.65%	79.37%	25.15%	76.62%	33.40%	77.99%
VideoMAE	Linear	47.54%	75.83%	13.36%	58.35%	30.45%	67.09%
ActionCLIP	Linear	35.71%	81.94%	33.05%	80.33%	34.38%	81.14%
DINOv2	KLD	*43.22%*	***88.80%***	*29.67%*	***88.61%***	*36.45%*	***88.71%***
VideoMAE	KLD	***47.74%***	*78.19%*	*20.04%*	*76.42%*	*33.89%*	*77.31%*
ActionCLIP	KLD	33.93%	79.96%	33.68%	80.33%	33.81%	80.15%

Table 1 presents the results of the three foundation models on our validation set. The results of the systems are substantially different between the two tasks. While VideoMAE and DINOv2 consistently outperform ActionCLIP on the action recognition task with top-1 accuracies of 47.74% and 43.22% compared to 40.48%, ActionCLIP excels in action anticipation, achieving a top-1 accuracy of 35.15%. On this task, VideoMAE and DINOv2 only yield 20.04% and 29.67% at

best. As expected, these results also indicate that recognizing the current action is generally a simpler task compared to anticipating the next action. Extending the output scope from one prediction to five predictions (top-5 accuracy), DINOv2 produces the best results across both tasks with 88.80% and 88,61%. This is at least 6.1% better than the other two models. Investigating the effect of the three proposed loss weighting strategies shows that VideoMAE and DINOv2 profit from the additional KL Divergence loss of *KLD*, while ActionCLIP performs best when trained with the vanilla equal weighting. The linearly adjusted loss amounts to no measurable improvements.

4.2 Results on the Test Set

Table 2: Results of the three foundation models on the official test set for the T3 Challenge 2025. Top-1 accuracies are reported separately for recognition, anticipation, and the combination. Bold font highlights the best overall results.

	Recognition	Anticipation	Total
Model	Top-1	Top-1	Top-1
DINOv2	**30.50%**	**30.73%**	**30.61%**
VideoMAE	30.24%	29.11%	29.69%
ActionCLIP	25.03%	24.66%	24.85%

We report the results of DINOv2, VideoMAE, and ActionCLIP on the official test set of the T3 Challenge 2025 in Tab. 2. We selected these model configurations due to the strong performance on the challenge's open development test set. The results in Tab. 2 indicate that all three models have difficulties generalizing to the new data, especially for the action recognition task. For action anticipation, the results have improved slightly for DINOv2 and VideoMAE, compared to our validation set. It is also noticeable that the results for the two tasks are roughly at the same level, and for DINOv2, the results on anticipation are even slightly better than for recognition. Surprisingly, ActionCLIP yields the worst performance on this test set, although it produced the best and most balanced combined accuracy scores in Table 1 using the validation set.

Overall, DINOv2 slightly outperforms VideoMAE in all three metrics, with an overall top-1 accuracy of 30.61%. However, the difference between the two methods is very small at only 0.9% on average. The poor performance of ActionCLIP on the 2025 T3 test set suggests that it could not fully utilize its zero-shot capabilities. Note that we limited the training to a relatively small number of epochs due to time constraints. We expect slightly improved results, with longer training schedules.

Investigating the individual results in more detail reveals that all models likely struggle with the same videos. In general, only a few videos are classified with moderate accuracy for both recognition and anticipation. Instead, the performance on all clips of a video tends to be either strong or poor. This could

be a result of the test set composition, as it contains two types of videos: 36 unseen regular procedure videos and 43 just-in-time procedure videos. It could also reflect frequently occurring classes in the training set as being learning relatively well, whereas rare edge cases, including many classes with only one or two samples, are hardly learned at all.

5 Conclusion

We adapted and evaluated three widely known foundation models – DINOv2, VideoMAE, and ActionCLIP – on the recognition and anticipation of medical procedure actions as part of the T3 Challenge 2025. The models reflect three different methods of processing the video information: frame-based, video-based, and a multi-modal approach based on video and text. We first adapted the three models to the challenge task, investigating prompt handling, hyperparameters, and loss weighting. Our evaluation indicates that the three foundation models show a decent performance on the recognition task and perform strongly with multiple possible predictions (top-5 accuracy). On the official test set, DINOv2 yields the best results with an overall accuracy of 30.61%. Despite this limited overall accuracy on the test set, our evaluation indicates the potential of foundation models in this data-scarce medical application with as few as one or two samples per class.

References

1. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: ViViT: A video vision transformer. In: International Conference on Computer Vision (2021)
2. Awais, M., Naseer, M., Khan, S., Anwer, R.M., Cholakkal, H., Shah, M., Yang, M.H., Khan, F.S.: Foundation models defining a new era in vision: a survey and outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: International Conference on Computer Vision (2021)
4. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the Kinetics Dataset. In: Computer Vision and Pattern Recognition (2017)
5. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. In: International Conference on Learning Representations (2024)
6. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (2005)
7. Donahue, J., Hendricks, L.A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(4), 677–691 (2017). https://doi.org/10.1109/TPAMI.2016.2599174
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)

9. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: Computer Vision and Pattern Recognition (2016)
10. Gerhardt, R.T., Berry, J.A., Blackbourne, L.H.: Analysis of life-saving interventions performed by out-of-hospital combat medical personnel. Journal of Trauma and Acute Care Surgery **71**(1) (2011)
11. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Computer Vision and Pattern Recognition (2022)
12. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
13. Oakden-Rayner, L.: Exploring large-scale public medical image datasets. Academic radiology **27**(1) (2020)
14. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024)
15. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (2021)
16. Tong, Z., Song, Y., Wang, J., Wang, L.: VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In: Advances in Neural Information Processing Systems (2022)
17. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. In: International Conference on Computer Vision (2015)
18. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Gool, L.V.: Temporal segment networks for action recognition in videos. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(11) (2018)
19. Wang, M., Xing, J., Liu, Y.: ActionCLIP: A new paradigm for video action recognition (2021), `https://arxiv.org/abs/2109.08472`
20. Yuan, L., Gundavarapu, N.B., Zhao, L., Zhou, H., Cui, Y., Jiang, L., Yang, X., Jia, M., Weyand, T., Friedman, L., et al.: VideoGLUE: Video general understanding evaluation of foundation models. Transactions on Machine Learning Research (2023)
21. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: iBOT: Image bert pre-training with online tokenizer. In: International Conference on Learning Representations (2022)
22. Zhuo, Y., Wachs, J., Zhang, E., Yu, X., Pachpande, A., Kirkpatrick, A.W., McKee, J., Rojas-Muñoz, E., Vanvoorst, B., Kwak, J.T., Tiomkin, S., Bebarta, V., Gao, Y.: The trauma thompson challenge 2025. `https://t3challenge25.grand-challenge.org/`, last accessed 2025-09-11