Figures and figure supplements

Scratch-AID, a deep learning-based system for automatic detection of mouse scratching behavior with high accuracy

Huasheng Yu and Jingwei Xiong *et al.*

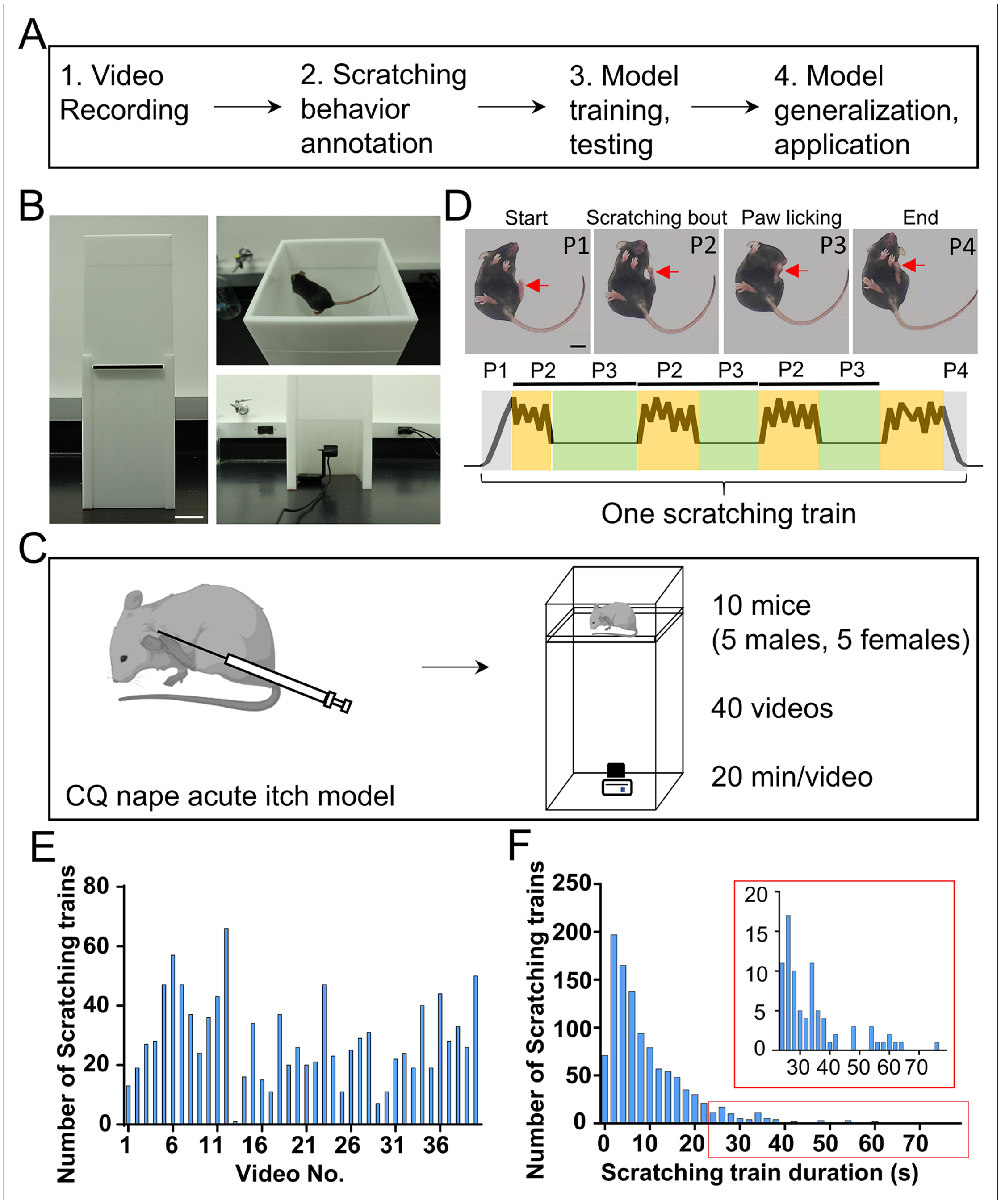


Figure 1. The overall workflow and building a customized videotaping box for mouse scratching behavior recording. (**A**) A diagram showing the workflow to develop a deep learning-based system for automatic detection and quantification of mouse scratching behavior. (**B**) An image of the designed videotaping box for high-quality video recording of mouse scratching behavior. Scale bar, 5 cm. (**C**) A cartoon showing the acute itch model induced by the chloroquine (CQ) injection in the nape, followed by video recording in the customized videotaping box. (**D**) Representative images showing different phases (P1–P4) of a scratching train (upper). Red arrows indicate the scratching hind paw. A cartoon showing the dynamic movement of the scratching hind paw in a scratching train (bottom). The cycle of scratching bout (P2) and paw licking (P3) may repeat once or more times in a scratching train. Scale bar, 1 cm. (**E**) The total number of scratching trains in each video. (**F**) The distribution of scratching train duration (*n* = 1135 scratching trains). The inset is the zoom-in of the red rectangle.

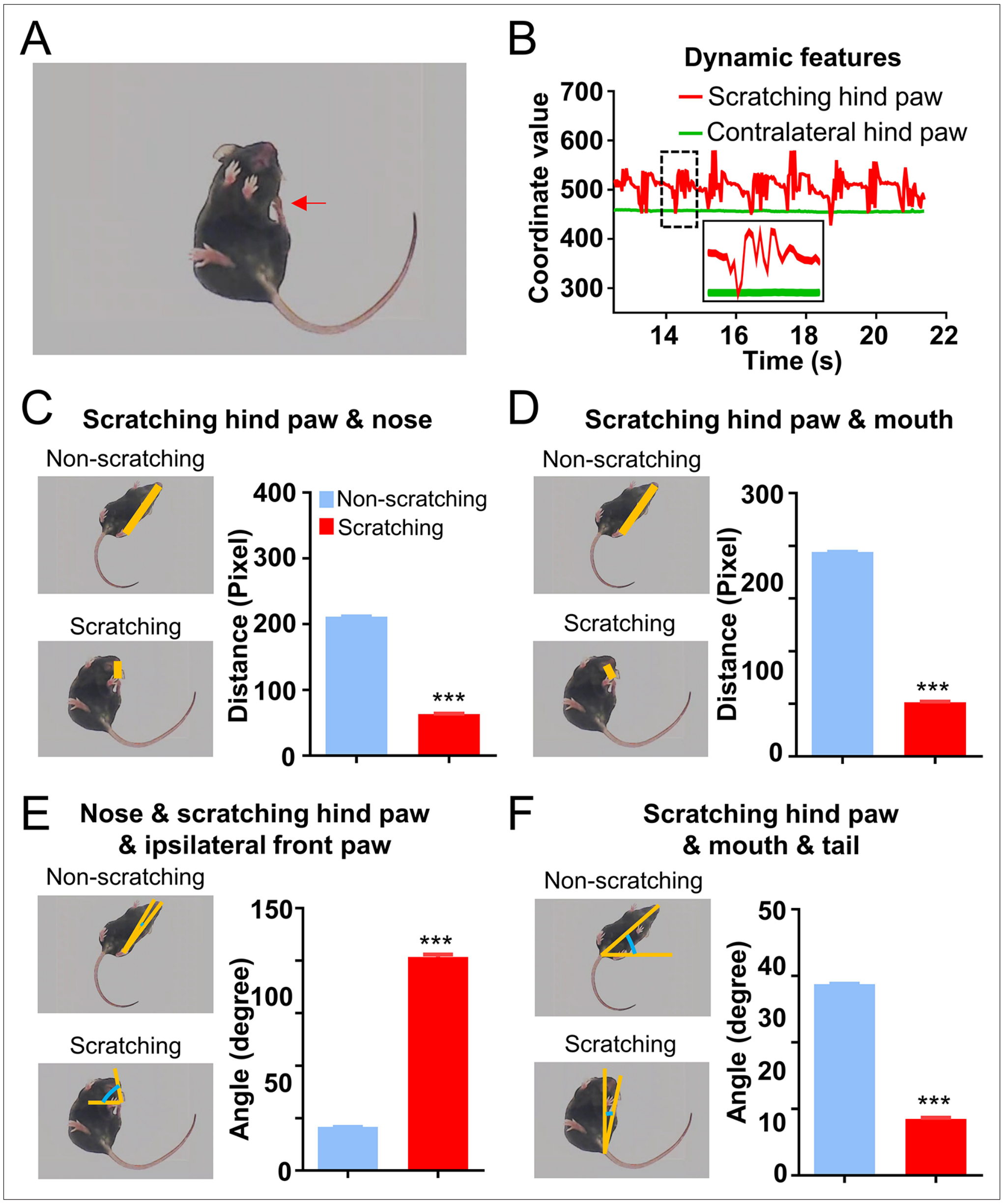
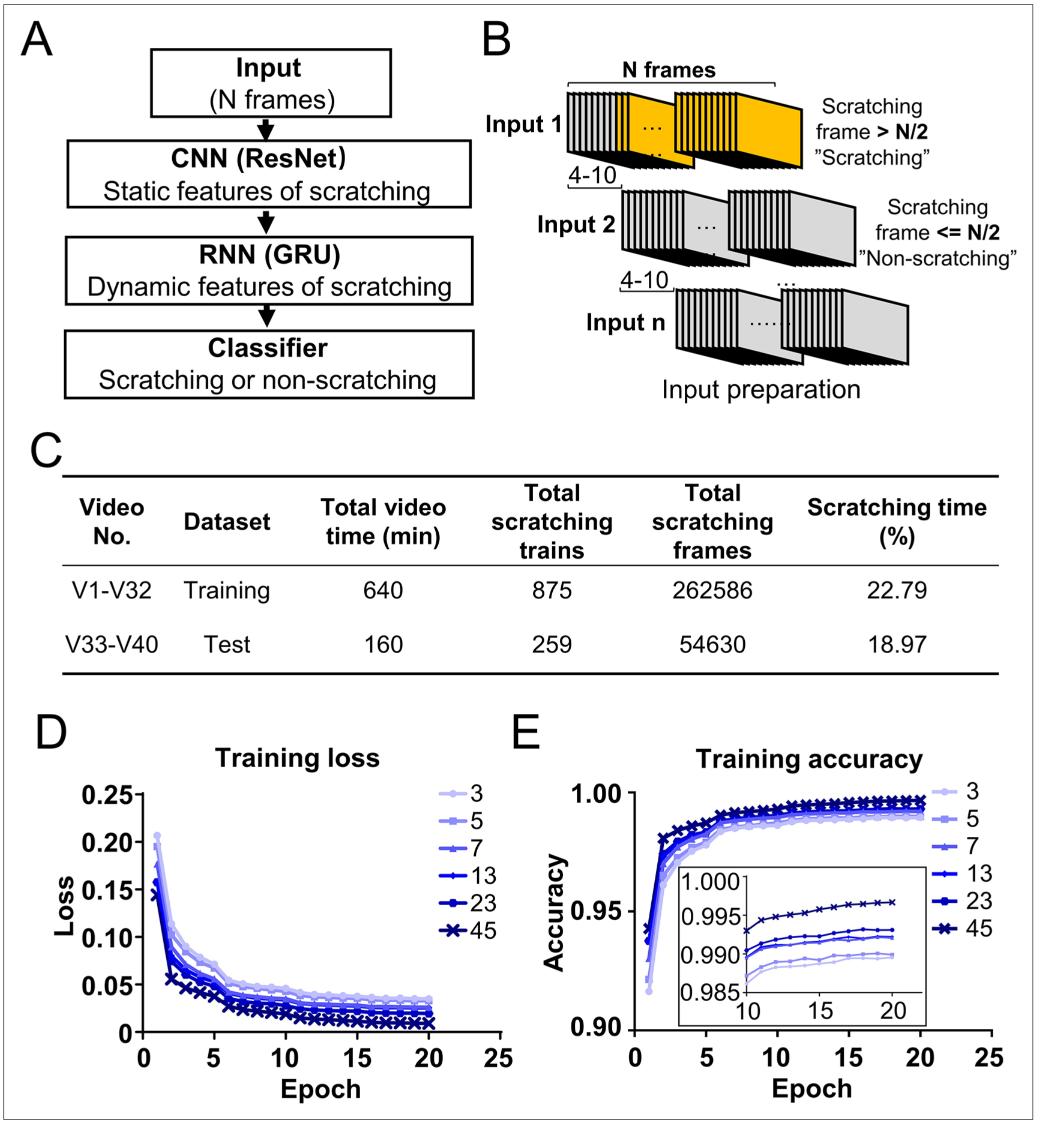


Figure 1—figure supplement 1. Dynamic and static features of scratching behavior in the chloroquine (CQ) nape acute itch model. (**A, B**) The scratching hind paw (red arrow in A) but not the contralateral hind paw displayed rhythmic vibration during scratching behavior (**B**). The distance between scratching hind paw and nose (**C**) or mouth (**D**), the angle consisting of nose, scratching hind paw, and ipsilateral front paw (**E**), and the angle consisting of scratching hind paw, mouth, and tail (**F**) in the scratching or non-scratching frames (*N* = 1756 frames for non-scratching, 2214 frames for scratching). Error bar, standard error of the mean (SEM). Differences between the two groups were analyzed using unpaired two-tailed Student's *t*-test,*** p < 0.001.



Video No.	Dataset	Total video time (min)	Total scratching trains	Total scratching frames	Scratching time (%)
V1-V32	Training	640	875	262586	22.79
V33-V40	Test	160	259	54630	18.97

Figure 2. Deep learning neural network design and training. (**A**) Cartoon showing the architecture of designed deep learning neural network consisting of the combination of convolutional neural networks (CNN), recurrent neural networks (RNN), and classifier. (**B**) Cartoon showing the preparation of inputs for the training dataset. Consecutive *N* frames were selected as one input for training. The interval between two adjacent inputs in a video was 4–10 frames. (**C**) The information of a sample training and test datasets. The training loss decreased (**D**) while the accuracy increased (**E**) during the training process with different input length (*N* = 3, 5, 7, 13, 23, 45 frames). The inset is the zoom-in of part of the figure.

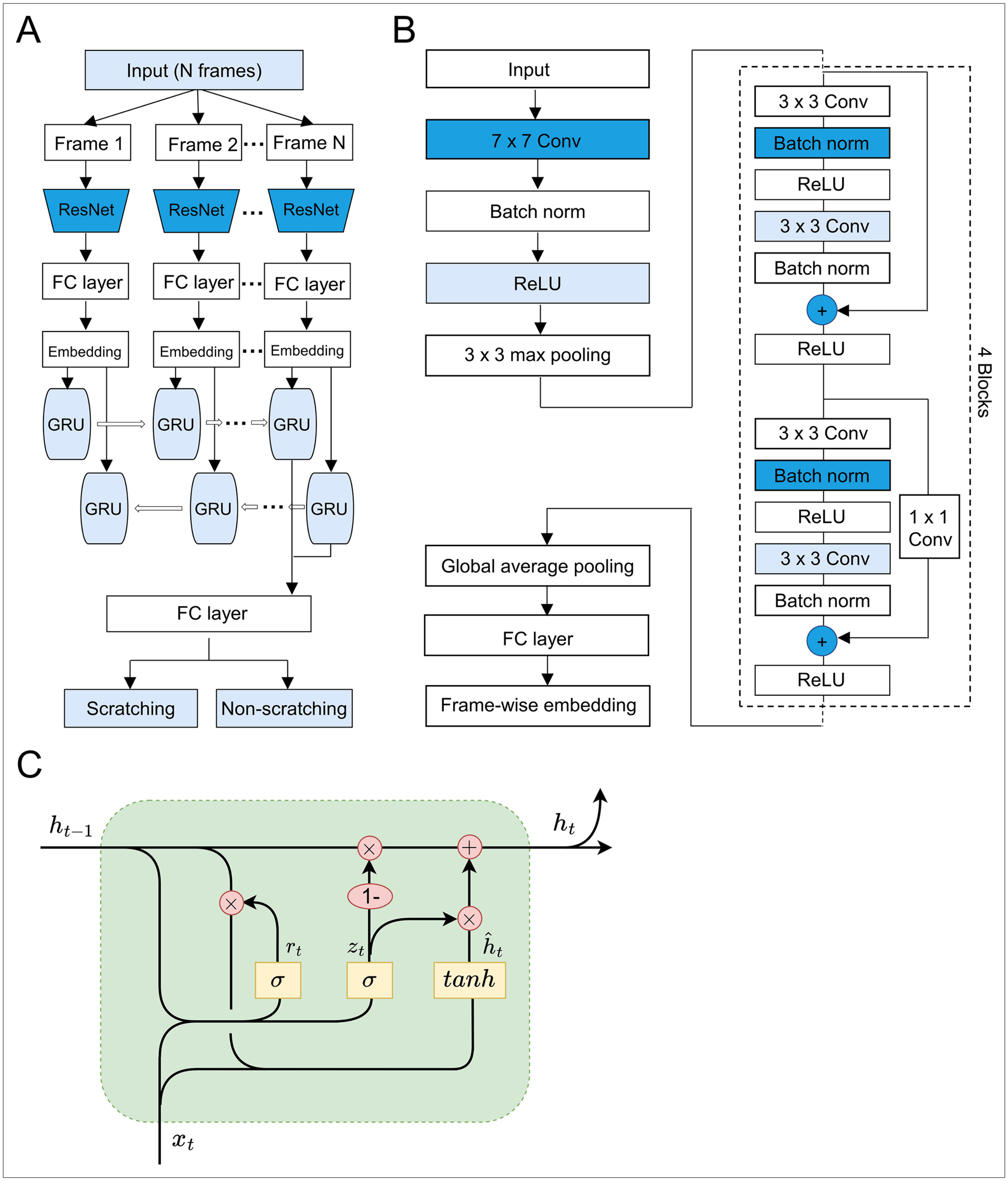


Figure 2—figure supplement 1. The architecture of deep learning neural network. (**A**) The cartoon showing the architecture of designed convolutional recurrent neural network (CRNN). The input was first fed into convolutional neural networks (CNN, ResNet-18), then followed by a two-layer bidirectional gated recurrent unit (GRU). A final full connection (FC) layer generated the binary prediction results. (**B**) Details of the modified ResNet-18 network. The final FC layer was modified to output frame-wise embedding instead of classification result. Conv, convolutional layer; Batch norm, batch normalization layer; ReLU, rectified linear unit; '+', element wise plus for vector. (**C**) Details of one GRU of the two-layer bidirectional GRU. The unit connected to embedding of frame *t*(x_t) took the last unit's output (h_{t-1}) as input and generated new output (h_t). The three yellow squares inside the unit were the reset gate, update gate, and candidate activation vector. 'x', Hadamard product for vectors; '+', vector plus.

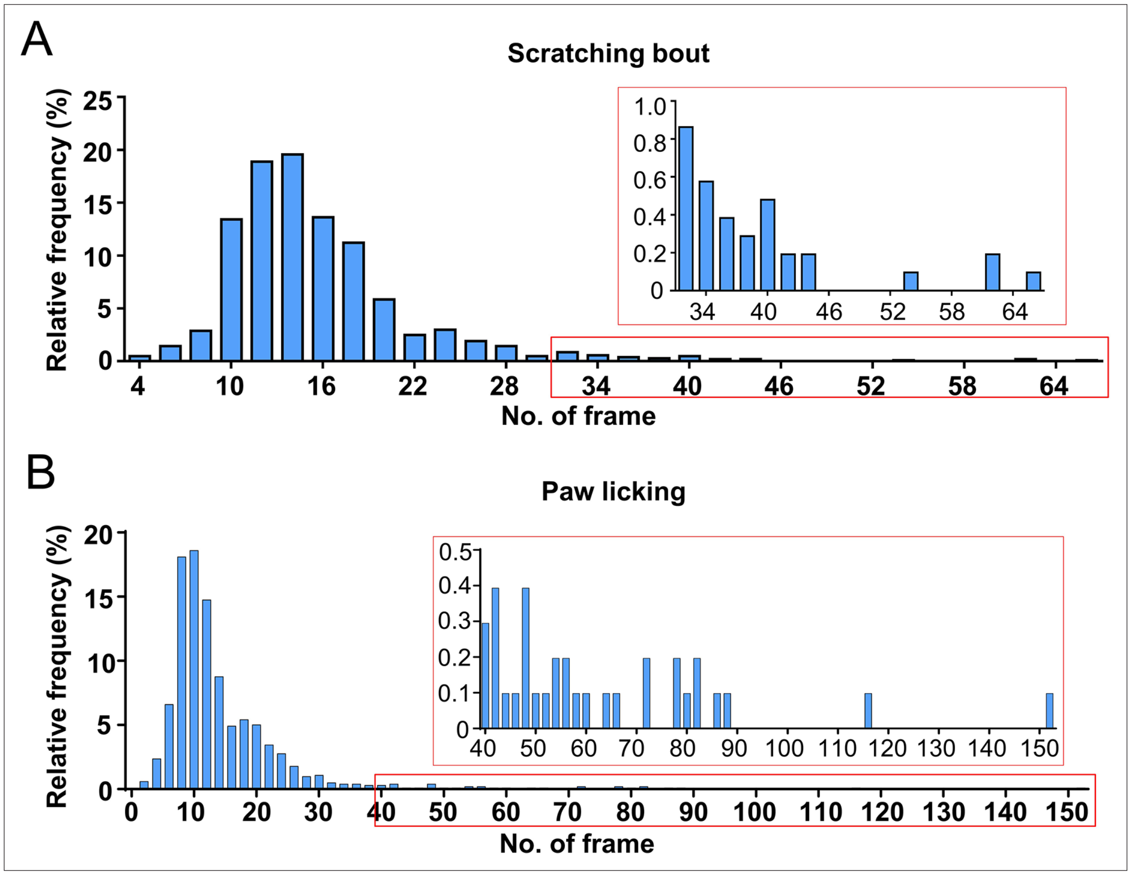


Figure 2—figure supplement 2. Distribution of durations of scratching bout and paw licking in the chloroquine (CQ) nape acute itch model. The frequency distribution of scratching bout (**A**) or paw licking (**B**) durations. The inset figure is the zoom-in of the red square part.

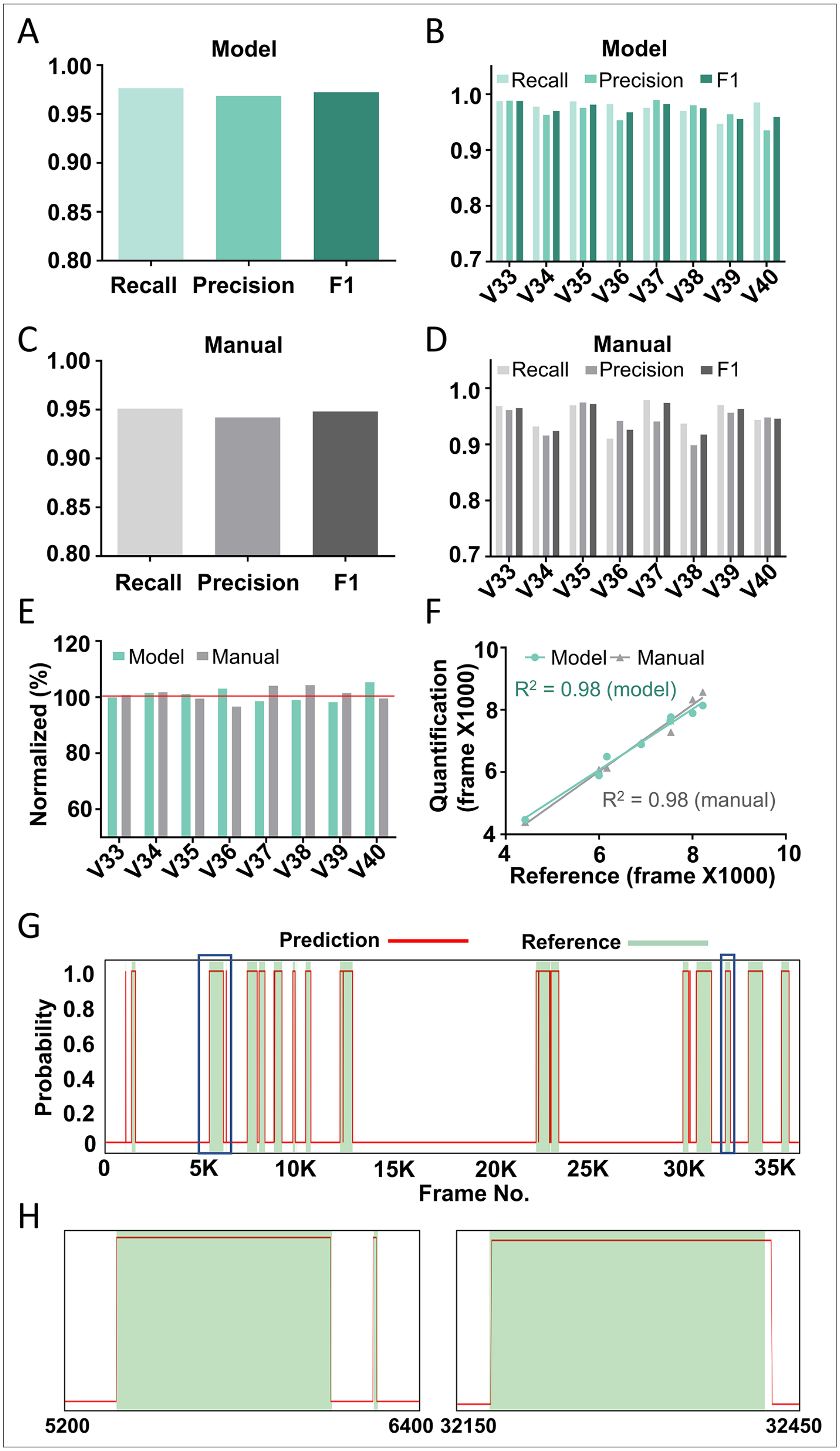


Figure 3. Performance of the best model on test videos. The recall, precision, and *F*1 score of the best model on average (**A**) or in individual videos (**B**). The recall, precision, and *F*1 score of manual annotation on average (**C**) or in individual videos (**D**). (**E**) The comparison among model prediction, manual quantification, and the reference annotation. The reference annotation is normalized to 100% shown as the red line. (**F**) The correlations between

Figure 3 continued on next page

Figure 3 continued

model prediction or manual quantification and reference annotation. R^2, Pearson correlation coefficient. (**G**) An example scratching probability trace (red curve) predicted by the model and aligned with the reference annotation (green bar). (**H**) The two zoom-ins from (**G**) showing the nice alignment between the model prediction and the reference annotation.

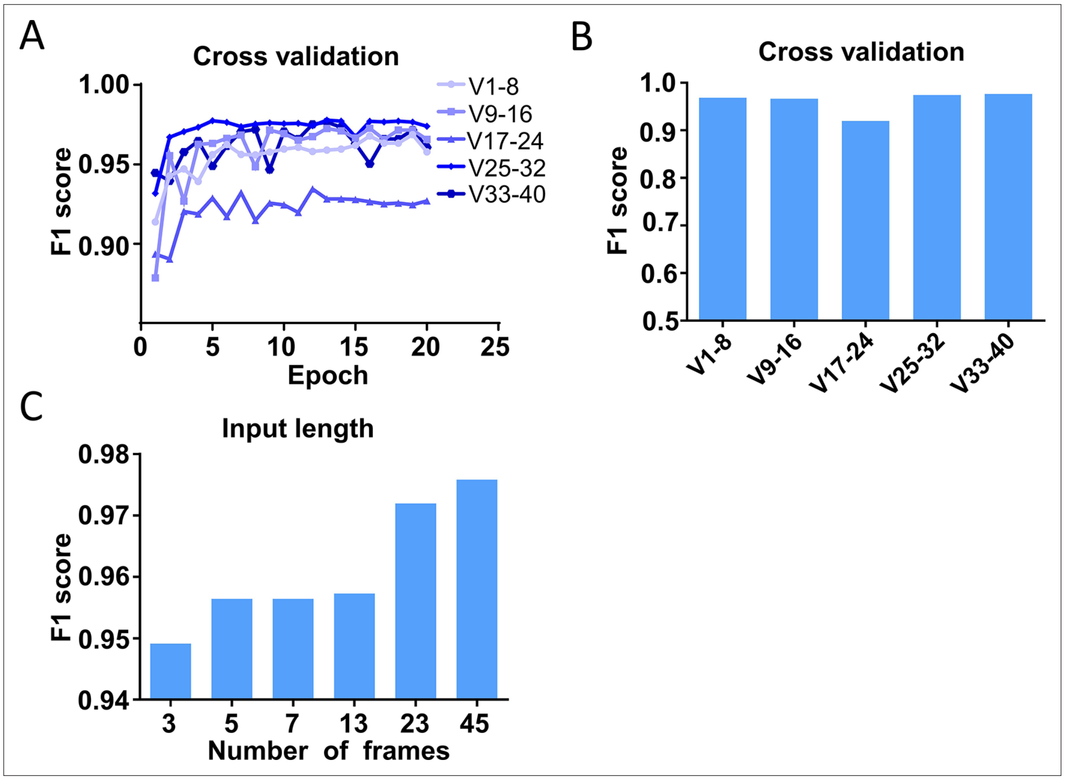


Figure 3—figure supplement 1. Cross-validation and parameter optimization of the prediction models. (**A**) Cross-validation of the trained prediction models by rotating the training and test datasets in 40 videos. The *F*1 score was calculated on the eight test videos. (**B**) The best performance of the models trained with input length N=45 and different combos of training and test datasets. (**C**) Model performances with different input lengths.

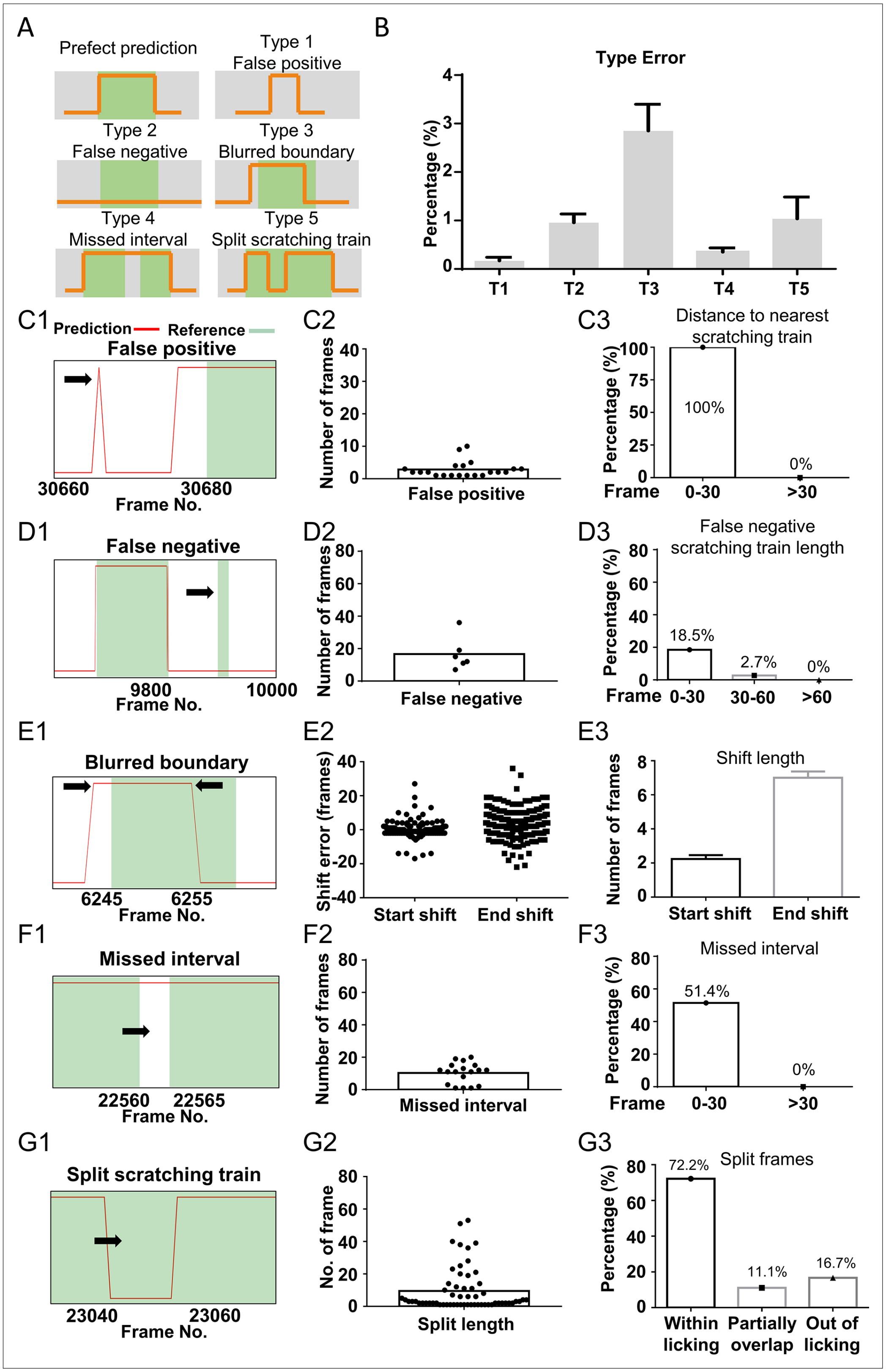


Figure 3—figure supplement 2. Error analysis of the best prediction model. (**A**) Cartoon showing five types of prediction errors. Red curves indicated scratching probability predicted by the model, and green bars indicated reference scratching trains. (**B**) The incidence rate of each type of errors, calculated by the ratio of the total frames in each type error over the total scratching frames of 8 test videos. Error bar, standard error of the mean (SEM).

Figure 3—figure supplement 2 continued on next page

Figure 3—figure supplement 2 continued

A real example of false positive prediction (**C1**), the duration of all false positive scratching trains (**C2**), and the frequency distribution of distances between false positive scratching trains to the nearest real scratching train (**C3**). A real example of false negative prediction (**D1**), the duration of all false negative scratching trains (**D2**), and the Type 2 error rate for scratching trains with different duration (**D3**). A real example of blurred boundary prediction (**E1**), the distribution (**E2**), and the average lengths (**E3**) of start and end shift. Error bar, SEM. A real example of missed interval (**F1**), the duration of all missed intervals (**F2**), and Type 4 error rate for intervals with different duration (**F3**). A real example of split scratching train (**G1**), the length of split frames (**G2**), and distribution of the Type 5 error linked to paw licking (**G3**).

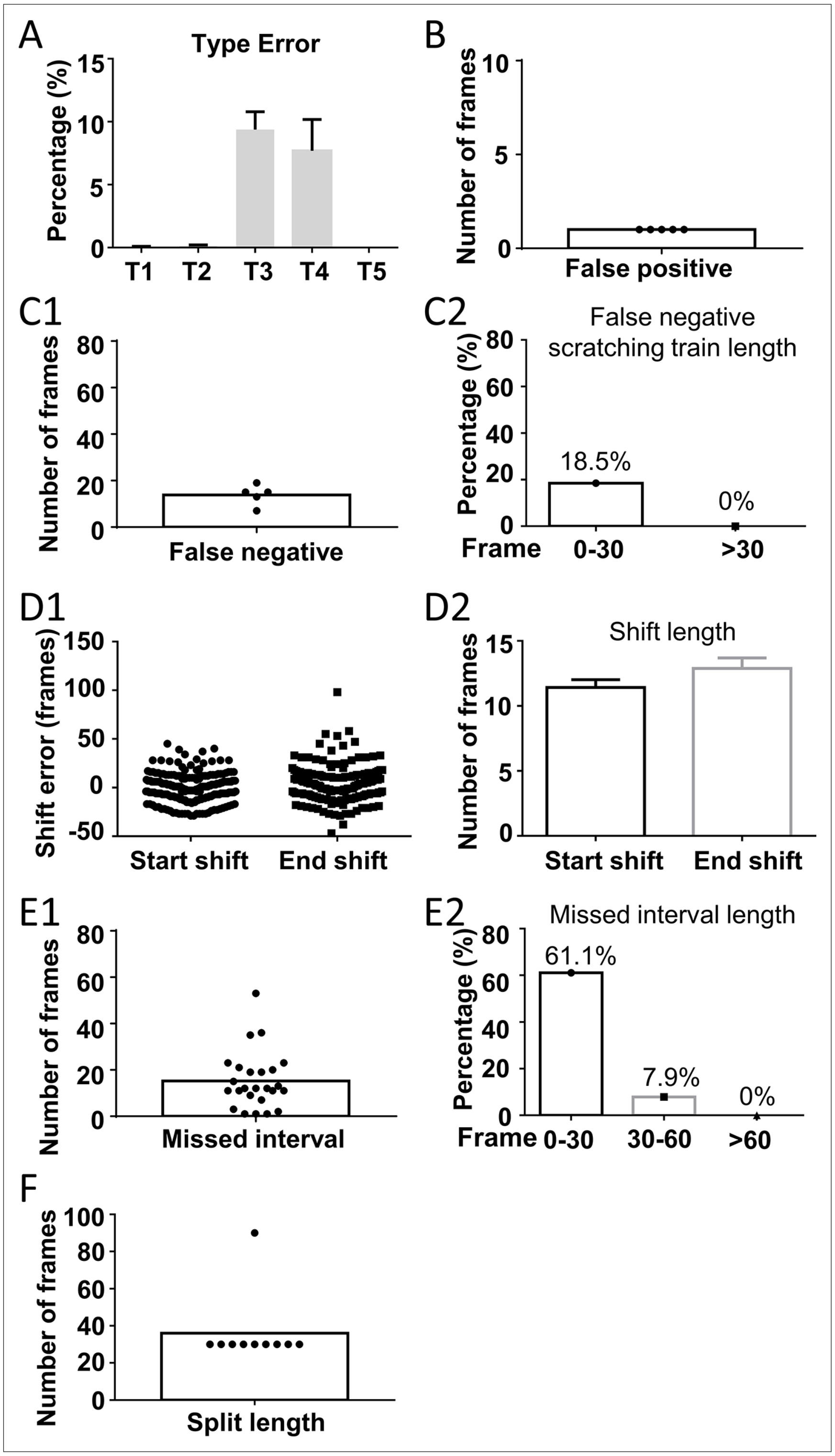


Figure 3—figure supplement 3. Error analysis of the manual quantification. (**A**) The incidence rate of each type of errors in manual annotation, calculated by the ratio of the total frames in each type error over the total scratching frames of eight test videos. Error bar, standard error of the mean (SEM). (**B**) The duration of all false positive scratching trains. The duration of all false negative trains (**C1**) and the Type 2 error rate for scratching trains with different duration (**C2**). The distribution (**D1**) and the average lengths (**D2**) of start and end shift. The duration of all missed intervals (**E1**) and the Type 4 error rate for intervals with different duration (**E2**). (**F**) The length of split frames.

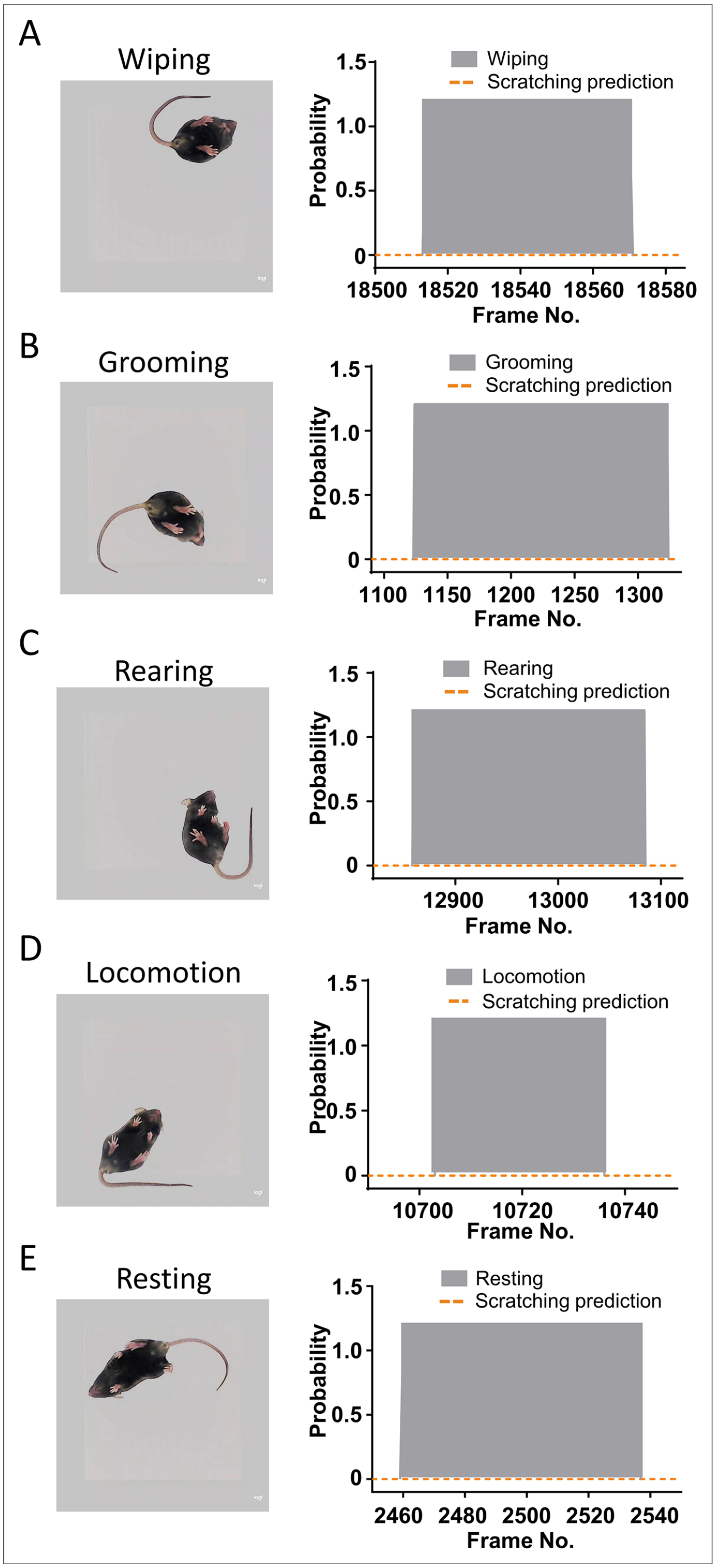


Figure 3—figure supplement 4. Other mouse behaviors were not recognized as scratching behavior. Examples of scratching behavior prediction probability during wiping (**A**), grooming (**B**), rearing (**C**), locomotion (**D**), and resting (**E**).

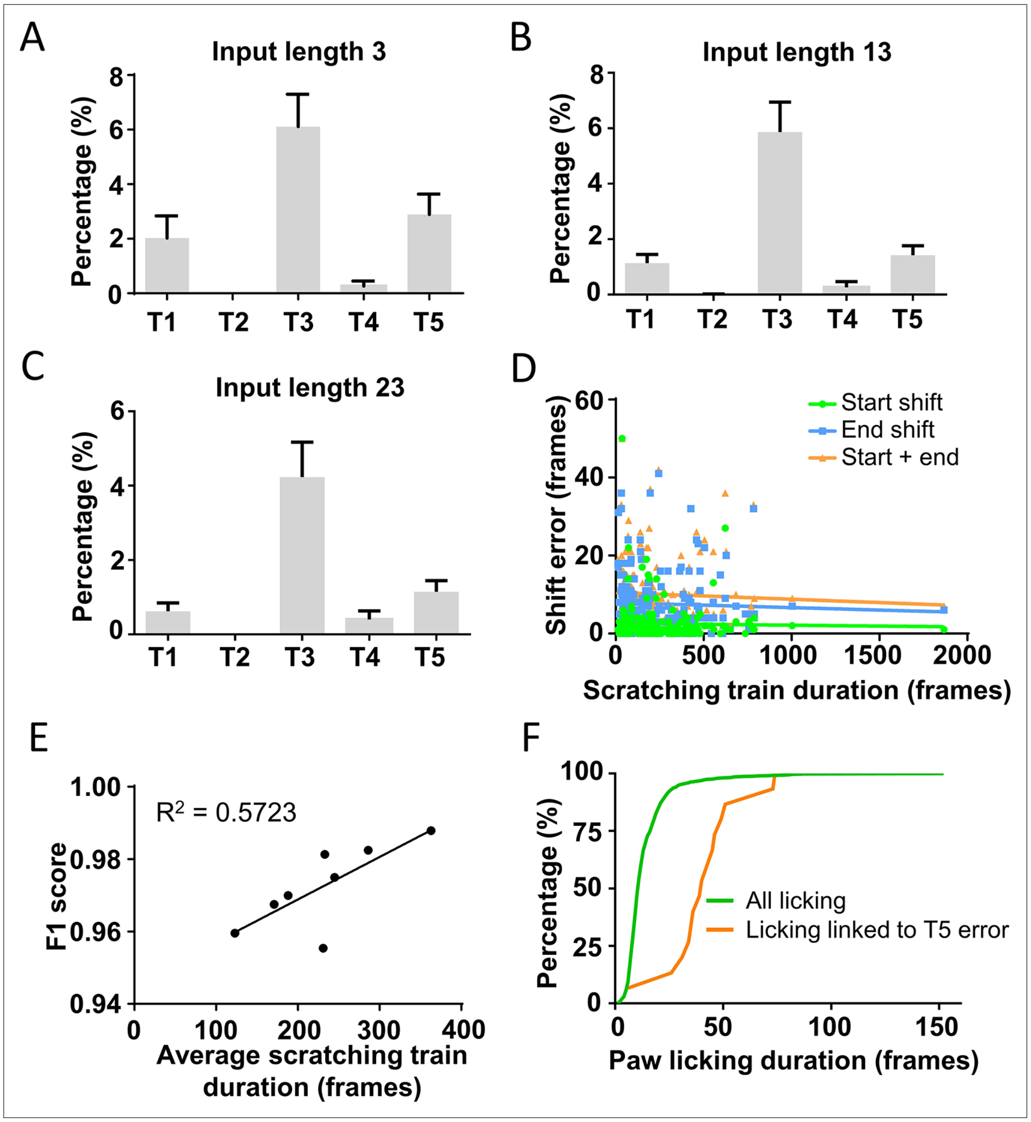


Figure 3—figure supplement 5. Relationship between prediction errors and the input length or the scratching train duration. (**A–C**) The five type error rates of models trained with different input lengths. Error bar, standard error of the mean (SEM). (**D**) The correlation between the scratching train duration and the length of start shift, end shift, or start shift plus end shift. (**E**) The correlation between the average scratching train duration in a video and the prediction accuracy (*F*1 score). (**F**) Frequency distribution of all paw licking duration and those linked to Type 5 error.

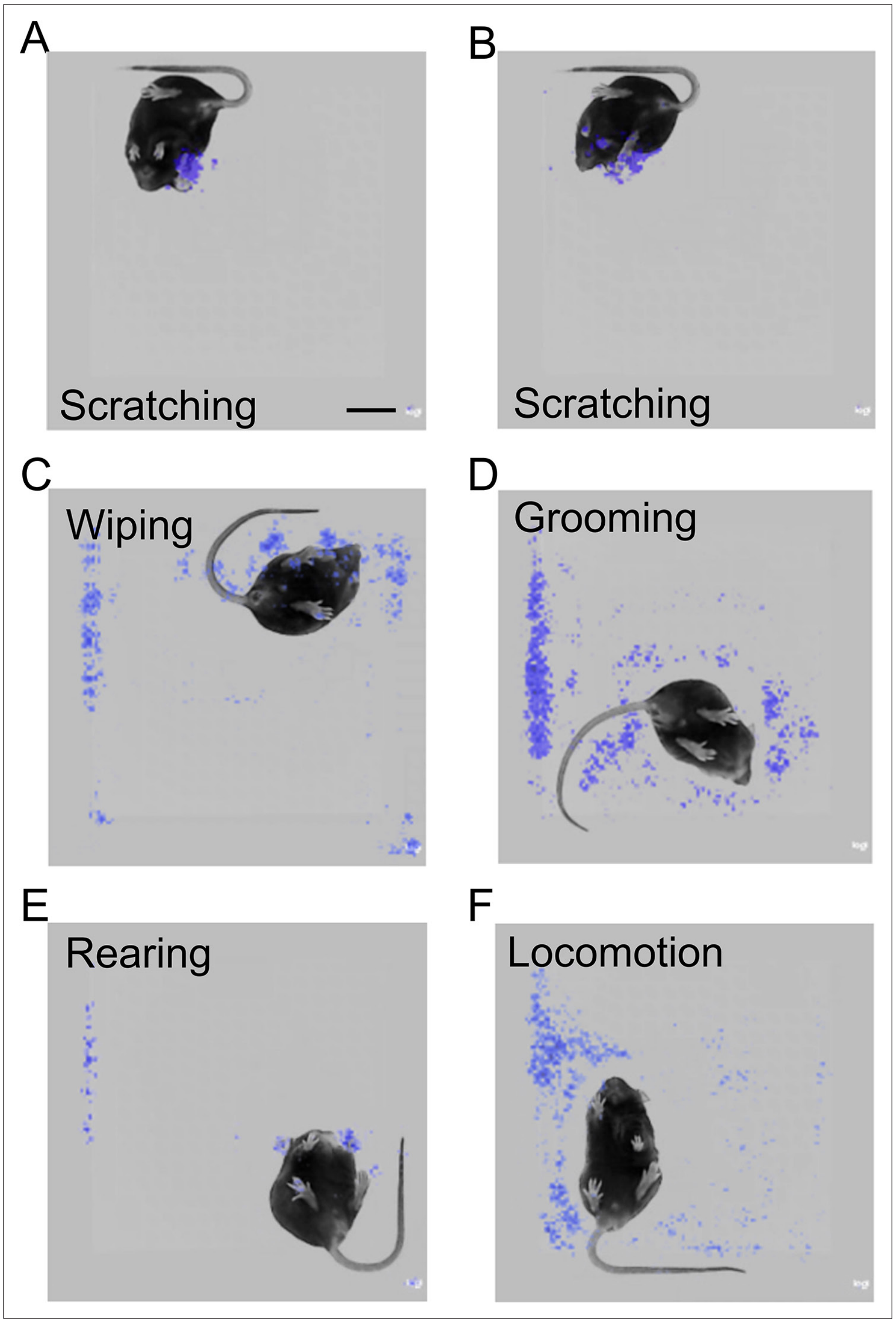


Figure 4. The prediction model focused on the scratching hind paw for scratching behavior recognition. (**A, B**) Saliency map showing the gradient value of each pixel of scratching frames during mouse scratching behavior prediction by the best model. The model focused on the scratching hind paw (**A, B**) and other body parts, such as front paws (**B**).Scale bar, 2 cm. Saliency map showing the gradient value of each pixel of wiping (**C**), grooming (**D**), rearing (**E**), and locomotion (**F**) frames during mouse scratching behavior prediction by the model.

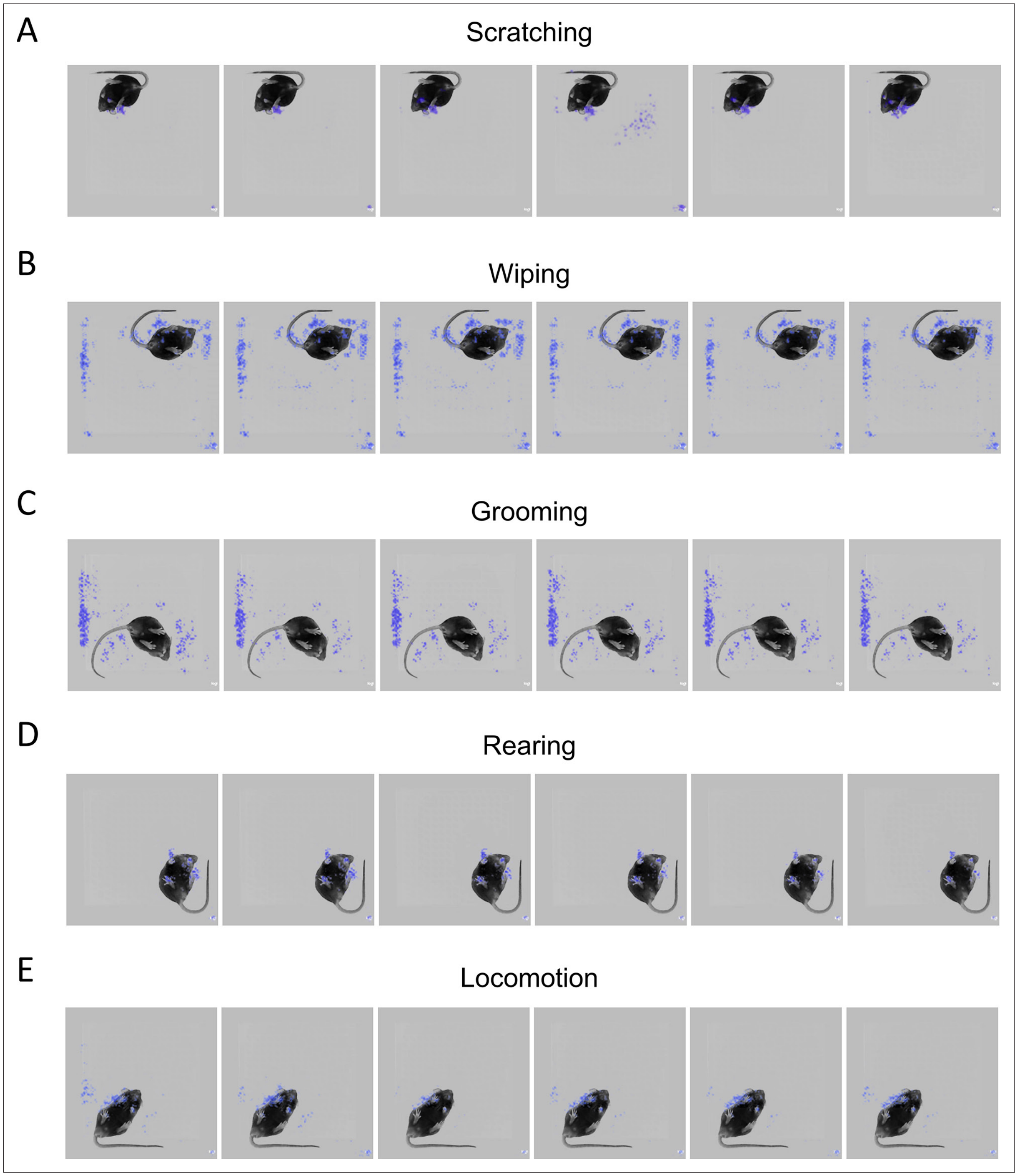


Figure 4—figure supplement 1. Saliency map of mouse scratching and other behaviors during the prediction. Additional saliency maps showing the gradient value of each pixel for frames of scratching (**A**), wiping (**B**), grooming (**C**), rearing (**D**), and locomotion (**E**) during the scratching behavior prediction of the best model.

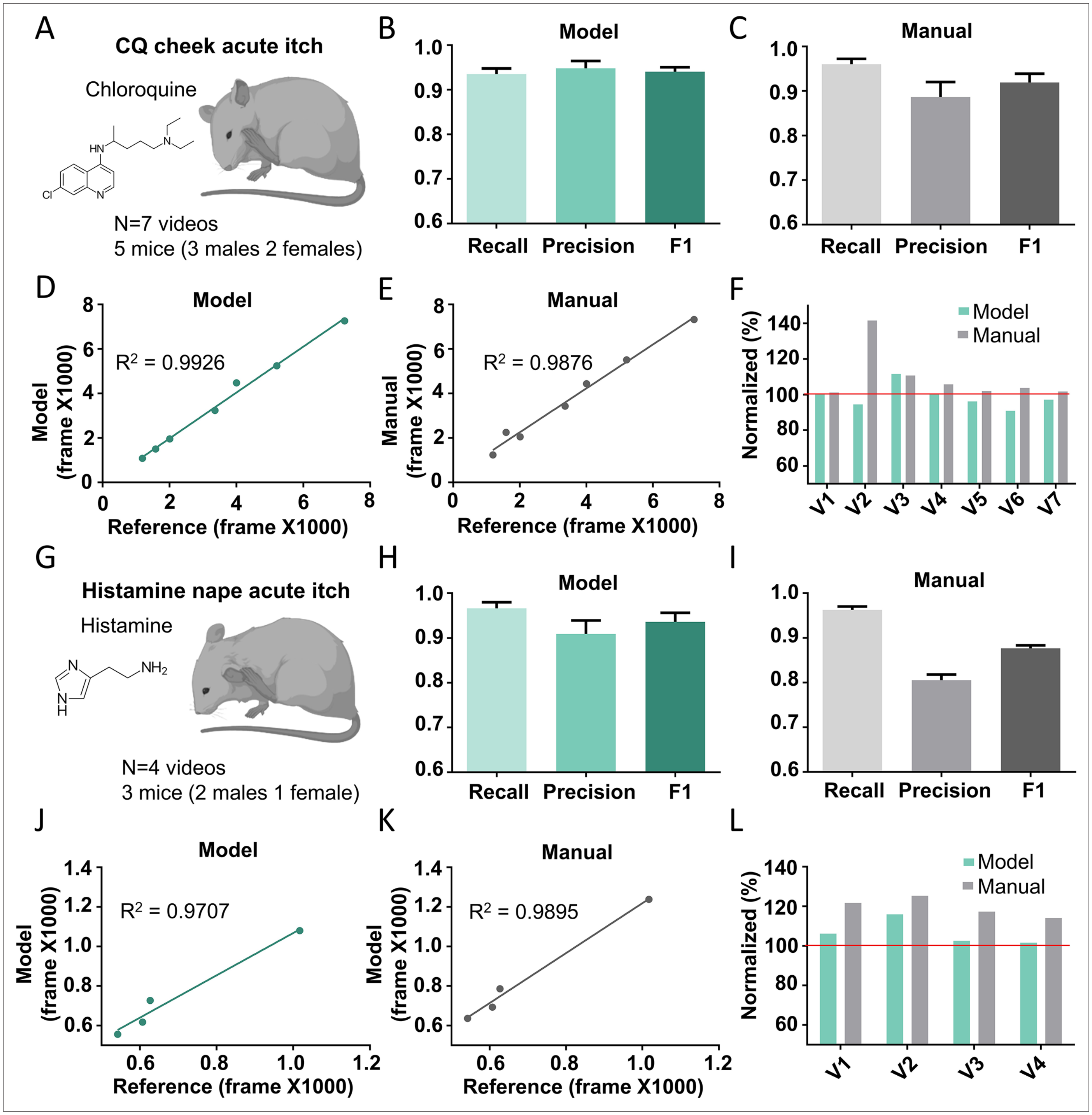


Figure 5. The Scratch-AID (Automatic Itch Detection) performance on other acute itch models. (**A**) A cartoon showing an acute itch model induced by chloroquine (CQ) injection in the mouse cheek. The average recall, precision, and *F1* score of Scratch-AID (**B**) or manual annotation (**C**). Error bar, standard error of the mean (SEM). The correlation between model prediction (**D**) or manual quantification (**E**) and reference annotation. R^2, Pearson correlation coefficient. (**F**) The comparison among model prediction, manual quantification, and reference annotation. The reference annotation is normalized to 100% shown as the red line. (**G**) A cartoon showing an acute itch model induced by histamine injection in the mouse nape. The average recall, precision and *F1* score of Scratch-AID (**H**) or manual annotation (**I**). Error bar, SEM. The correlation between model prediction (**J**) or manual quantification (**K**) and reference annotation. R^2, Pearson correlation coefficient. (**L**) The comparison among model prediction, manual quantification, and reference annotation. The reference annotation is normalized to 100% shown as the red line.

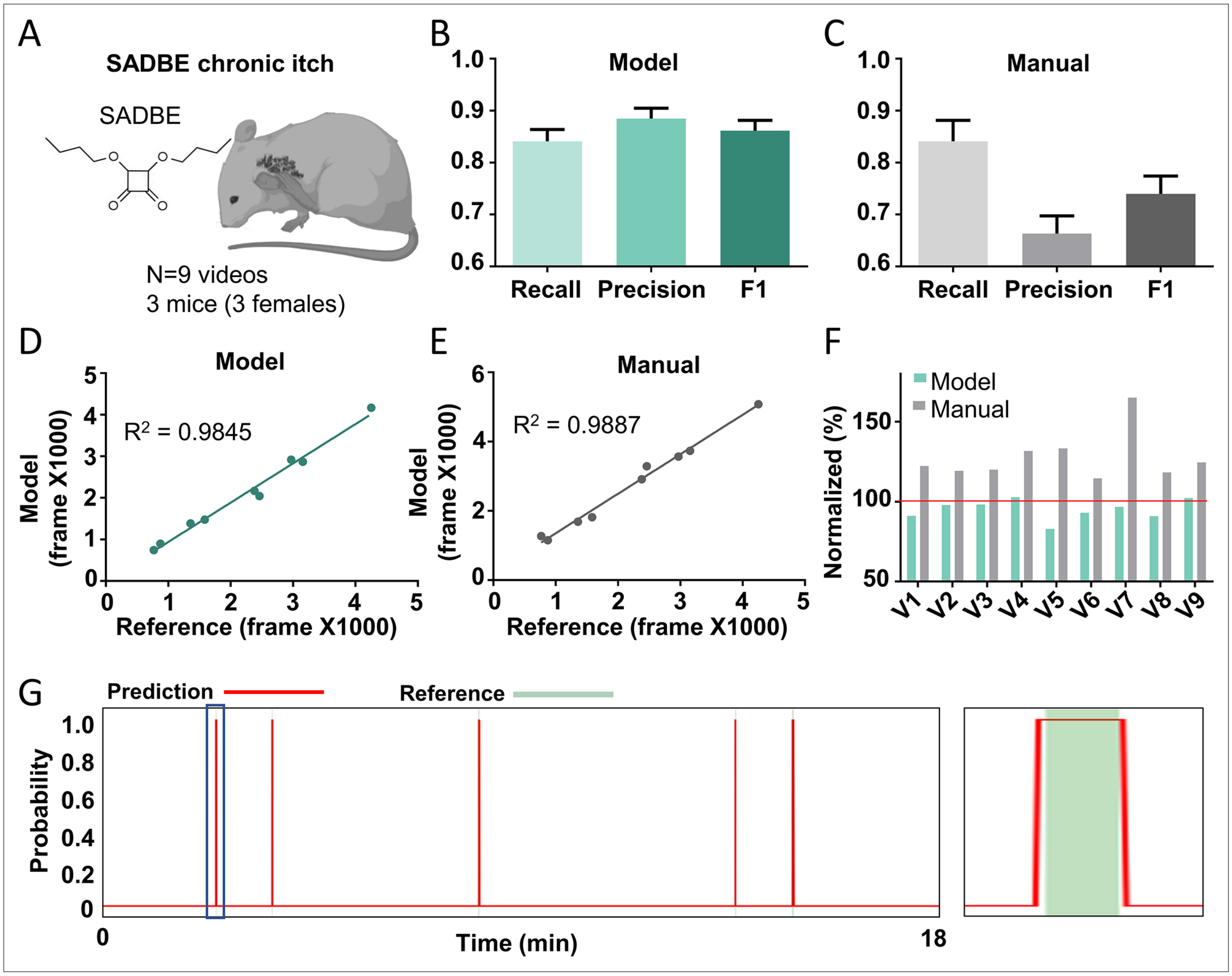


Figure 6. The Scratch-AID performance on a chronic itch model. (**A**) A cartoon showing a squaric acid dibutylester (SADBE) induced chronic itch model. The average recall, precision, and *F*1 score of Scratch-AID (**B**) or manual annotation (**C**). Error bar, standard error of the mean (SEM). The correlation between model prediction (**D**) or manual quantification (**E**) and reference annotation. R^2, Pearson correlation coefficient. (**F**) The comparison among model prediction, manual quantification, and reference annotation. The reference annotation is normalized to 100% shown as the red line. (**G**) An example scratching probability trace (red curve) predicted by the model and aligned with the reference annotation (green bar) (left). Zoom-in (right panel) of the blue square part showing nice alignment of the model prediction with the reference annotation.

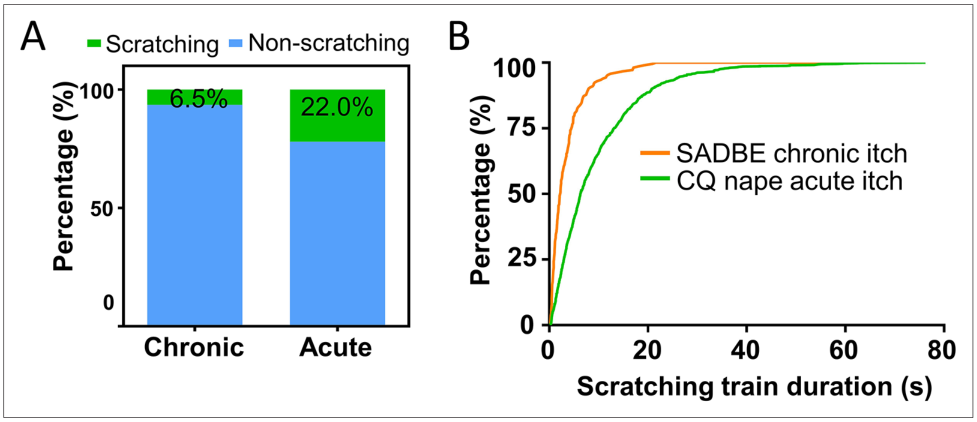


Figure 6—figure supplement 1. Different dynamic features of chronic and acute itch models. (**A**) The percentage of scratching and non-scratching frames in the squaric acid dibutylester (SADBE) chronic itch model (*n* = 9 videos) and chloroquine (CQ) nape acute itch model (*n* = 40 videos). (**B**) Frequency distribution of scratching train duration of SADBE chronic itch model and CQ nape acute itch model.

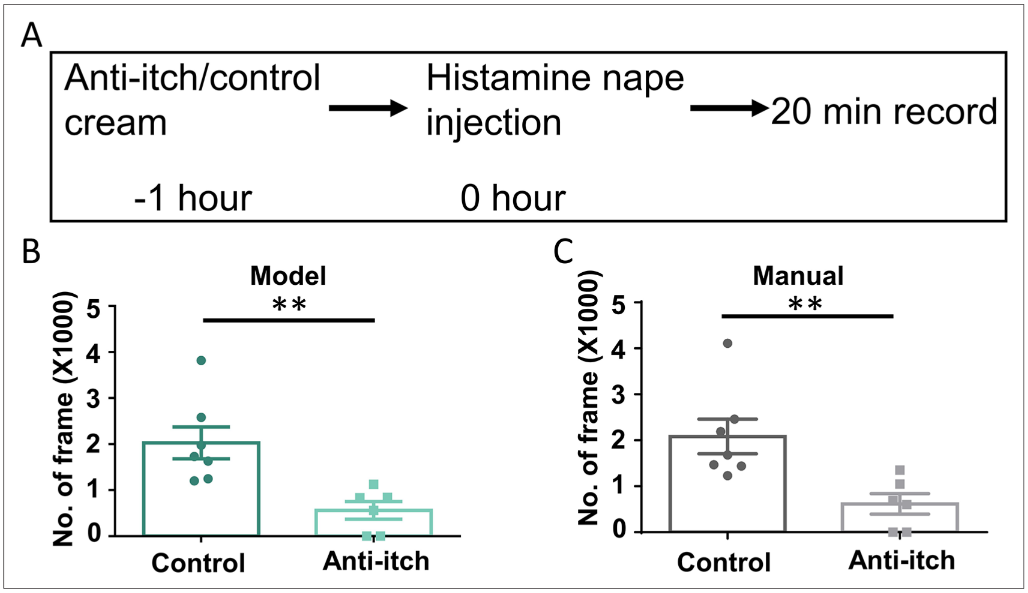


Figure 7. Application of the Scratch-AID (Automatic Itch Detection) system in a drug screening paradigm. (**A**) A diagram showing the experimental design of an anti-itch drug test. Quantification of scratching behavior in anti-itch cream treated group or control group by Scratch-AID (**B**) or manual annotation (**C**). Error bar, standard error of the mean (SEM). Differences between the two groups were analyzed using unpaired two-tailed Student's *t*-test, ** $p < 0.01$.