Supplementary Material for
VADER: Video Alignment Differencing and Retrieval

Alexander Black[1] Simon Jenni[2] Tu Bui[1] Md. Mehrab Tanjim[2] Stefano Petrangeli[2]
Ritwik Sinha[2] Viswanathan Swaminathan[2] John Collomosse[1,2]
[1]CVSSP, University of Surrey [2]Adobe Research
{alex.black,t.v.bui}@surrey.ac.uk {jenni,tanjim,petrange,risinha,vishy,collomos}@adobe.com

1. Video Demonstration

We include a video recording with a demonstration of VADER system end-to-end process in the supplementary materials. A screenshot of the demo is shown in Figure 1. Demo environment allows to choose a query video from ANAKIN and specify a type of manipulation (see Sec. 2). A top-1 full length original video is retrieved (Sec. 3.1 main paper) and aligned with the query (Sec. 3.2 main paper). The aligned candidate video and the query are passed through the differencing module (Sec. 3.3 main paper) and an outline of the edited region is shown as the final output.

2. ANAKIN

We provide examples and additional statistics about our presented dataset of mANipulated videos and mAsK annotatIoNs - ANAKIN. Figures 2, 3, 4 show duration distributions of the original videos, the trimmed-and-edited clips and rations between the two. The durations of full videos are spread quite uniformly between 5 and 240 seconds, while the mean length of the trimmed clips is $\sim$ 5 seconds. The average length ratio between the edited clips and original videos is 0.1.

We identify five types of manipulation tasks present in ANAKIN: splicing, inpainting, swap (color, background, text), frame-level manipulation (frame reversal, speed up/slow down, frame skipping) and audio manipulation (replacement, addition). We provide a detailed breakdown of the number of videos for each manipulation type in Table 1. Examples of the first three categories of manipulations are shown in Figure 6. These manipulations have an associated binary mask annotations provided, since the edits are localized within the frame. Examples of the latter two categories - audio and temporal - are shown in Figure 7. There are no binary mask annotations for these manipulation types.

3. Robust Video Descriptor Objectives

To train a robust video descriptor extraction model used in our retrieval module, we uniquely employ three self-supervised training objectives: visual and audio-visual contrastive terms as well as temporal reasoning tasks.

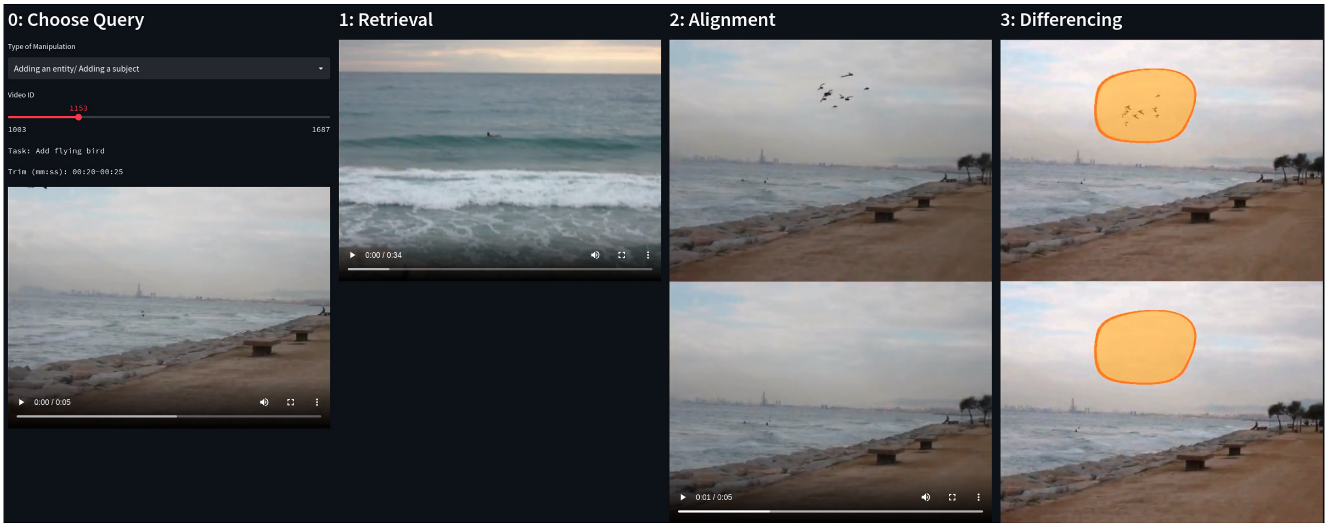
Contrastive Terms is adopted from the NTXentLoss [3]. Given a training video, positive visual contrastive learning pairs are constructed from differently augmented versions of clips from within the training instance. We use a wide variety of augmentations during training to ensure model robustness. Augmentations include space time cropping, temporal speed and direction manipulation, strong color jittering (hue, saturation, brightness, contrast, random grayscale, solarize), as well as random blur and flipping. Positive audio-visual pairs are constructed from the same clips and their corresponding temporally aligned audio tracks. In both cases, negatives are sampled from other video and audio clips from the current mini-batch or a memory bank of prior embeddings.

Temporal Reasoning include two sets of classification tasks: playback speed/direction and temporal ordering. For playback speed and direction classification, we employ a 8-way softmax loss: 4 speed (1,2,4,8x) levels x 2 forward/backward directions, applied to video and audio clips separately. For temporal ordering classification, we classify a random pair of temporal signals in both intra- and cross-modal fashion (video-video, audio-audio and video-audio) into 3 categories: (i) correctly ordered, (ii) overlapping, and (iii) wrongly ordered.

4. Alignment and Differencing Ablations

Alignment module is feature agnostic. Therefore, in addition to the performance achieved using RMAC [6] features shown in Tab. 4 of the main paper, we evaluate all of the methods using ICN [2] features in Table 2. Even though ICN features are designed to be more robust to benign transformations, all of the approaches, including ours, achieve lower scores of percentage alignment up to threshold.

Differencing module training is done in three stages. First, it is pre-trained for 88 epochs on PSBattles[2], an image dataset analogous to ANAKIN, where each image is repeated 16 times to match the video input size. Then, the model is fine-tuned on ANAKIN for 137 epochs to learn better temporal reasoning. Finally, we finetune our model for further 64 epochs with frame misalignment as an extra augmentation (using the same frame sampling strategy as described in Sec. 3.3 of the main paper). We report model's performance after each of the stages, comparing it to the



0: Choose Query

Type of Manipulation

Adding an entity/ Adding a subject

Video ID

1153

1003 1687

Task: Add flying bird

Trim (mm:ss): 00:20-00:25

1: Retrieval

2: Alignment

3: Differencing

Figure 1. Screenshot of the VADER demo platform. **0**: Filter ANAKIN videos by manipulation type and choose a specific video to use as a query. Textual description of the task that was given to the editor and the time from which the snippet was trimmed are shown for reference and are not visible to the model. **1**: Full length original video returned by the retrieval module is shown. **2**: The alignment module is used to localize the query within the retrieved video. Query (top) and corresponding original frames (bottom) are shown together. **3**: An outline of the manipulated region, produced by the differencing module, is shown on top of both the edited and the original videos.

Total: 1042									
Splicing: 252	Inpainting: 194	Swap: 261			Frame-level: 174			Audio: 161	
		Col: 111	BG: 54	Text: 92	Reverse: 66	Up/down: 103	Skip: 44	Add: 97	Replace: 64

Table 1. Breakdown of ANAKIN manipulation types.

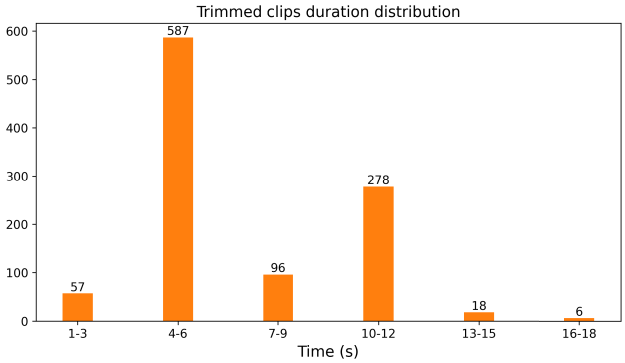


Figure 2. ANAKIN video statistics. Dduration distribution of the trimmed clips, from which the edited clips are made.

strongest single-frame baseline ICN [2]. Table 3 reports mean IOU scores for linear shifts between 0 and 3 frames. There is a significant improvement by 30% in VADER performance after training on ANAKIN. Introduction of temporal augmentations greatly improves the IOU scores in the cases with temporal shift present without affecting performance in case of perfect alignment. Figure 8 illustrates how VADER's resilience to temporal shifts increases after training on ANAKIN and additionally improves after the introduction of temporal augmentations in the training. Figure

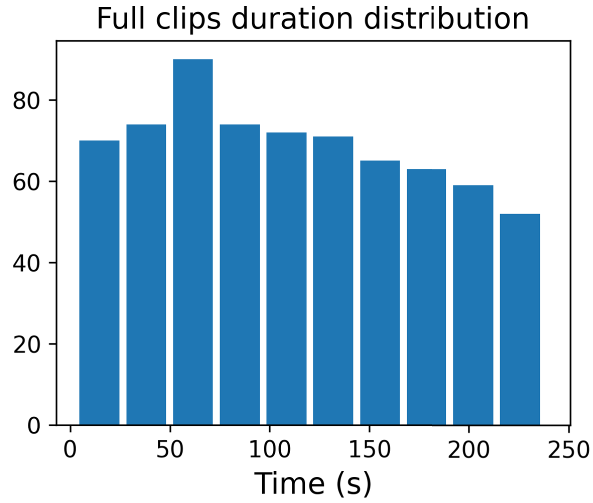


Figure 3. ANAKIN video statistics. Duration distribution of full length videos in the dataset.

5 shows a plot of VADER differencing module performance in IOU evaluated against subsets of ANAKIN of different sizes. VADER scores 0.801 ± 0.032 of IOU on 10% of ANAKIN test set, and 0.799 ± 0.002 at 95%, which indicates that the current dataset size is sufficient for evaluating

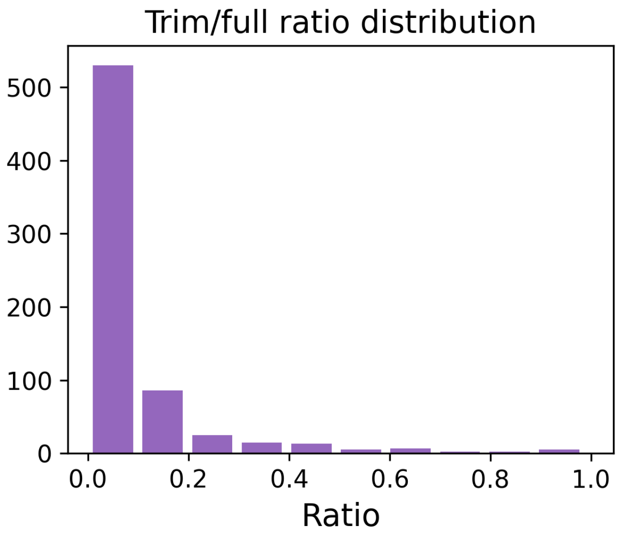


Figure 4. ANAKIN video statistics. Distribution of ratios of trimmed clips to full video lengths.

Method	**clean + benign**			**manip + benign**		
	@0.1s	**@1s**	**@10s**	**@0.1s**	**@1s**	**@10s**
			Features: RMAC [6]			
VADER	**64.2**	**78.4**	**93.2**	**53.7**	**74.2**	**91.6**
LAMV [1]	44.7	68.9	85.3	30.0	59.5	83.2
CTE [4]	4.2	20.1	67.2	3.7	19.6	68.3
TMK [5]	1.6	18.4	63.2	3.2	17.9	59.5
			Features: ICN [2]			
VADER	**48.4**	**57.4**	**82.1**	**32.6**	**44.2**	**74.7**
LAMV [1]	18.4	39.5	73.7	17.9	37.4	72.6
CTE [4]	2.6	15.3	68.3	2.1	15.3	67.7
TMK [5]	1.1	8.9	40.5	0.5	8.4	40.5

Table 2. Alignment performance using different backbone feature extractors.

Method, dataset	IOU			
Shift	0	1	2	3
VADER, ANAKIN+aug	0.804	**0.801**	**0.786**	**0.760**
VADER, ANAKIN	**0.808**	0.780	0.701	0.629
VADER, PSBattles	0.448	0.437	0.405	0.370
ICN [2]	0.448	0.408	0.372	0.347

Table 3. Evaluation of the differencing module at different training stages.

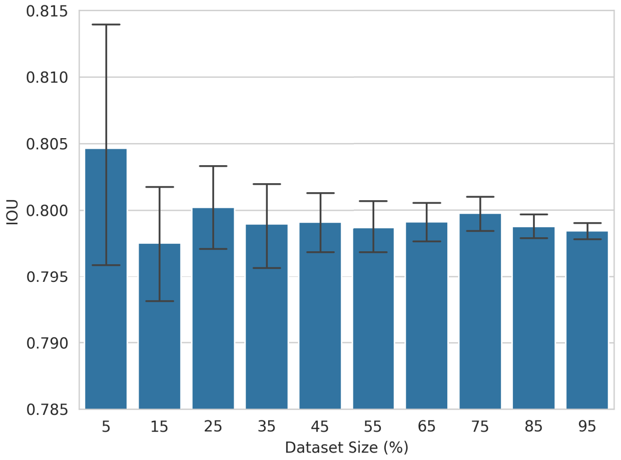


Figure 5. Differencing module performance (IOU) against ANAKIN test set size (%). Mean and confidence intervals from 100 samples.

the presented work.

References

[1] Lorenzo Baraldi, Matthijs Douze, Rita Cucchiara, and Hervé Jégou. Lamv: Learning to align and match videos with kernelized temporal layers. In *Proc. CVPR*, pages 7804–7813, 2018. 3

[2] Alexander Black, Tu Bui, Hailin Jin, Vishy Swaminathan, and John Collomosse. Deep image comparator: Learning to visualize editorial change. *Proc. CVPR WS*, 2021. 1, 2, 3

[3] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *Proc. ICML*, 2020. 1

[4] Matthijs Douze, Jérôme Revaud, Jakob J. Verbeek, Hervé Jégou, and Cordelia Schmid. Circulant temporal encoding for video retrieval and temporal alignment. *IJCV*, 119:291–306, 2015. 3

[5] Sébastien Poullot, Shunsuke Tsukatani, Anh Phuong Nguyen, Hervé Jégou, and Shin'ichi Satoh. Temporal Matching Kernel with Explicit Feature Maps. In *ACM Multimedia 2018*, pages 1–10, Brisbane, Australia, Oct. 2015. ACM Press. 3

[6] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. Particular object retrieval with integral max-pooling of cnn activations. *CoRR*, abs/1511.05879, 2016. 1, 3

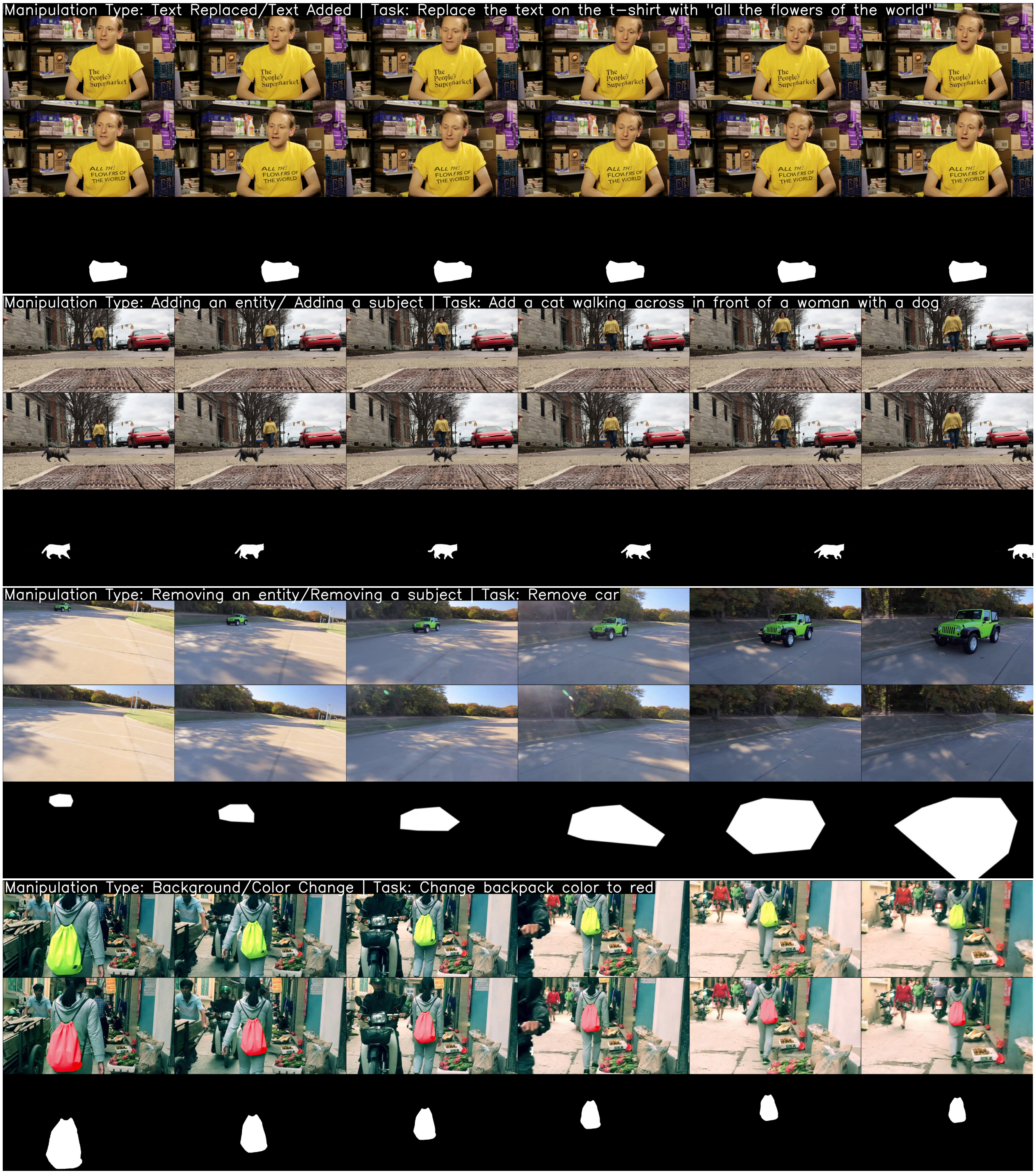

Manipulation Type: Text Replaced/Text Added | Task: Replace the text on the t−shirt with "all the flowers of the world"
Manipulation Type: Adding an entity/ Adding a subject | Task: Add a cat walking across in front of a woman with a dog
Manipulation Type: Removing an entity/Removing a subject | Task: Remove car
Manipulation Type: Background/Color Change | Task: Change backpack color to red

Figure 6. ANAKIN dataset examples of the videos with manipulations that are paired with binary mask annotations. Top-left corner of each example contains details about manipulation type and specific task given to the editor. For each of the four examples, original video, manipulated video and binary mask annotation are placed in top, mid and bottom rows, respectively.

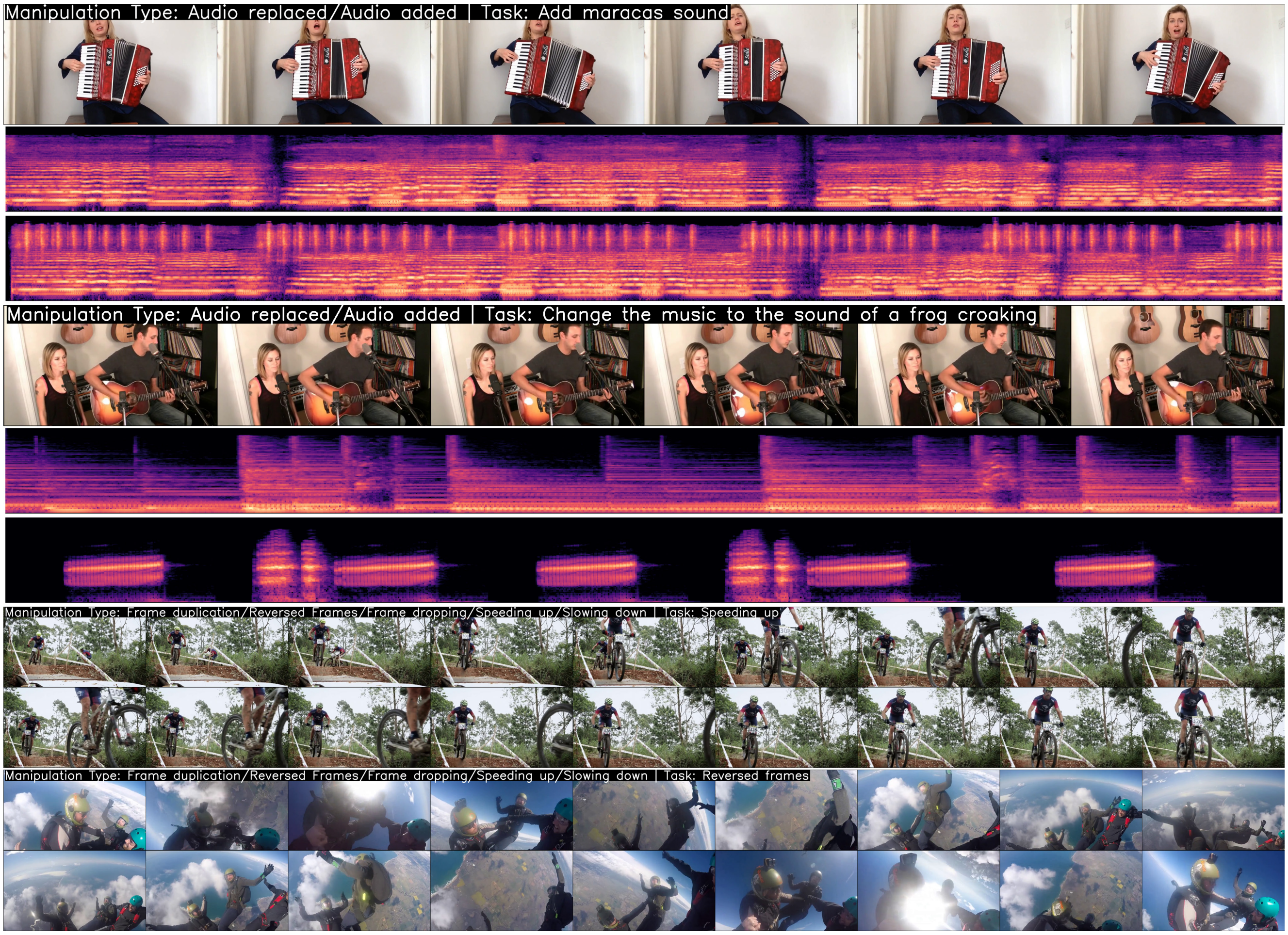

Manipulation Type: Audio replaced/Audio added | Task: Add maracas sound
Manipulation Type: Audio replaced/Audio added | Task: Change the music to the sound of a frog croaking
Manipulation Type: Frame duplication/Reversed Frames/Frame dropping/Speeding up/Slowing down | Task: Speeding up
Manipulation Type: Frame duplication/Reversed Frames/Frame dropping/Speeding up/Slowing down | Task: Reversed frames

Figure 7. ANAKIN dataset examples of the manipulations that are not paired with binary masks. First two examples: video, original- and manipulated audio. Final two examples: original video and frame-level manipulations.

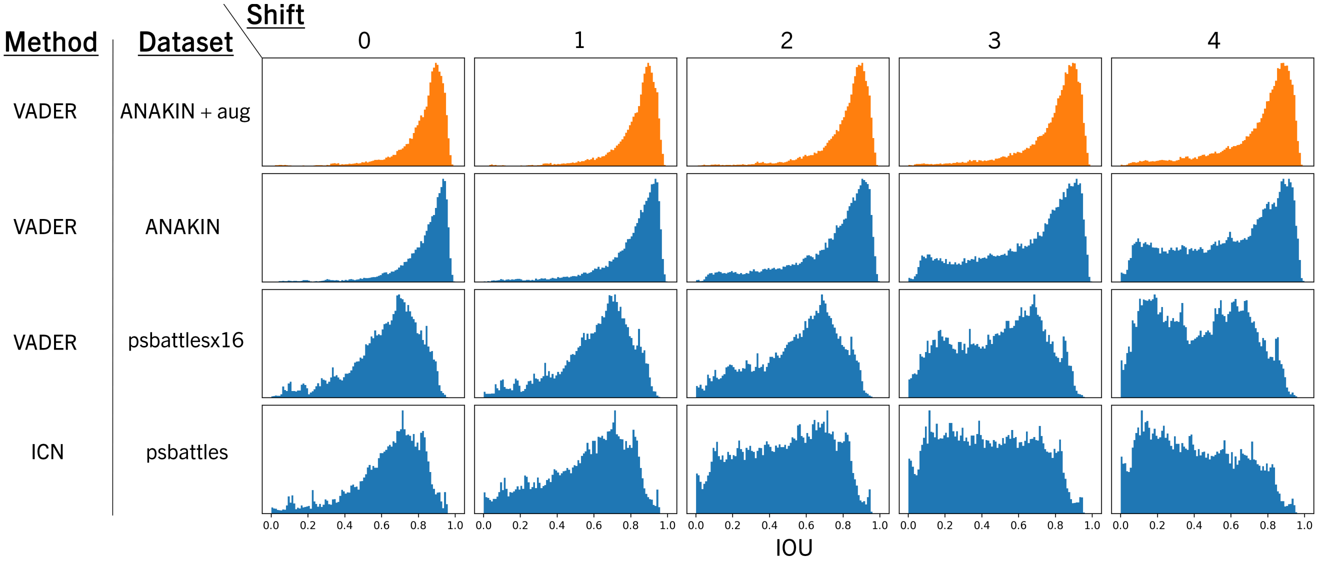


Figure 8. Distributions of IOU scores for different stages of VADER training, evaluated at different linear shifts between two videos.