A Large-scale Comprehensive Dataset and Copy-overlap Aware Evaluation Protocol for Segment-level Video Copy Detection: Supplementary Material

S1. Frame-to-frame similarity map

The frame-to-frame similarity matrix based on frame-level features is capable of representing temporal structure between the copied videos and it has proved to be successful to learn similarity patterns of the pairwise frame similarities [1, 2].

The matrix element s_{ij} is the similarity score between a frame feature f_i of video A and f'_j of a potential copied video B.

$$s_{ij} = \frac{f_i * f'_j}{||f_i|| * ||f'_j||} \tag{1}$$

where i,j indicates the ith frame of video A and the jth frame of the video B. An example of frame-to-frame similarity is shown in Figure 1. The red bounding box indicated to temporal copied part of two original videos is a copied segment pair. This pattern in the bounding box which is similar to an oblique straight line represents the temporal sequential copy between two videos with high similarity scores between the temporal corresponding frames.

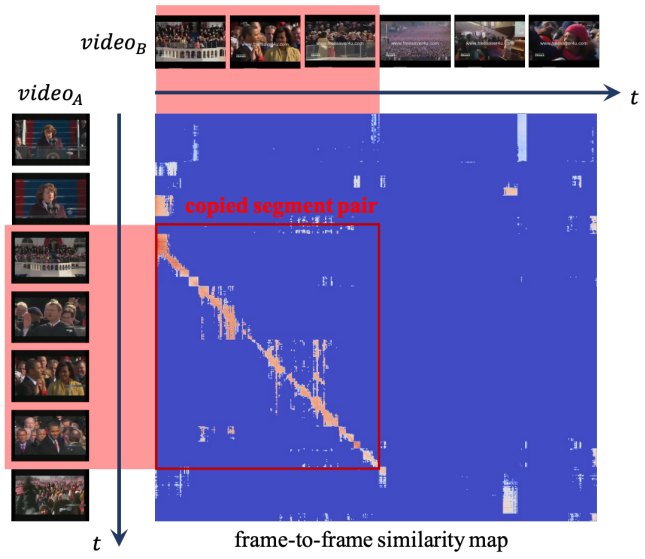

Figure 1. An frame-to-frame similarity map example.

S2. Example copied videos in VCSL

As we mentioned in Section 3.2 in our main text, VCSL covers lots of realistic spatial and temporal transformations. Due to the representation restriction by only figures and text descriptions in this manuscript, we only list some frame transformations (spatial transformation) here. We recommend readers to discover the various temporal transformations in VCSL covering reverse, loop, video mushup, acceleration and deceleration. Moreover, we have also tried to label different types of transformations on VCSL. But both the spatial and temporal transformations are too various to be concluded by limited types, and we finally give up this plan.

Fig.2 below shows some typical spatial transformations in VCSL including crop, filter, text overlay, background, cam-cording, picture in picture, even recent deepfake, etc. The videos marked with the same background blue box are from the same query set, and the first video inside each background box is the seed video. There are a wide range of content transformations among over 280k segment copies in VCSL, and these realistic skillful transformations bring great challenges to segment-level copy detection.

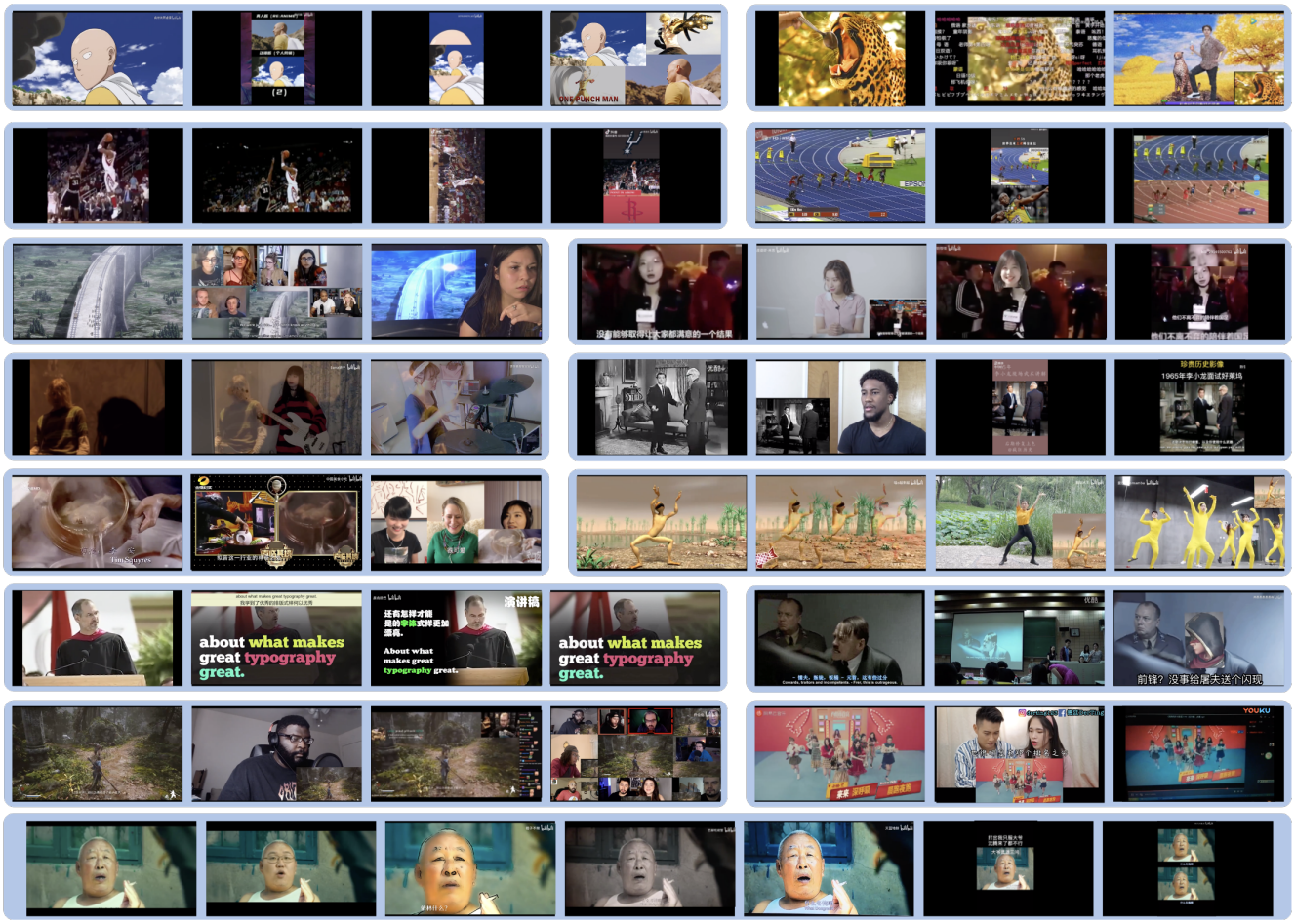

Figure 2. Example copied videos from VCSL. The videos with the same background blue box are from the same query set, and the first video of each background box is seed video.

S3. Similarity maps on hard cases

In Section 5.3 of our main text, we observe that the effect of features on the final results is not as dramatic as expected, especially considering different feature dimensions. By observing the similarity map, it can be found that the patterns of these copied segments are not obvious and show no contrast with the surroundings for some hard cases. Similarity maps on some example hard cases including picture in picture, severe crop and cam-cording are given in Fig.3.

The red bounding boxes in Fig.3 indicate the manually annotated copied locations. Theoretically, there should be distinct copied patterns at the ground truth copied locations (e.g., pattern shown in Fig.1) if the frame feature extractor works ideally. However, the unrecognizable copied patterns in Fig.3 show the limitation of current global features that they cannot extract the similarity between the temporal corresponding frames. We hope that the hard cases in VCSL give some insights to develop more powerful feature representations for the segment-level video copy detection task.

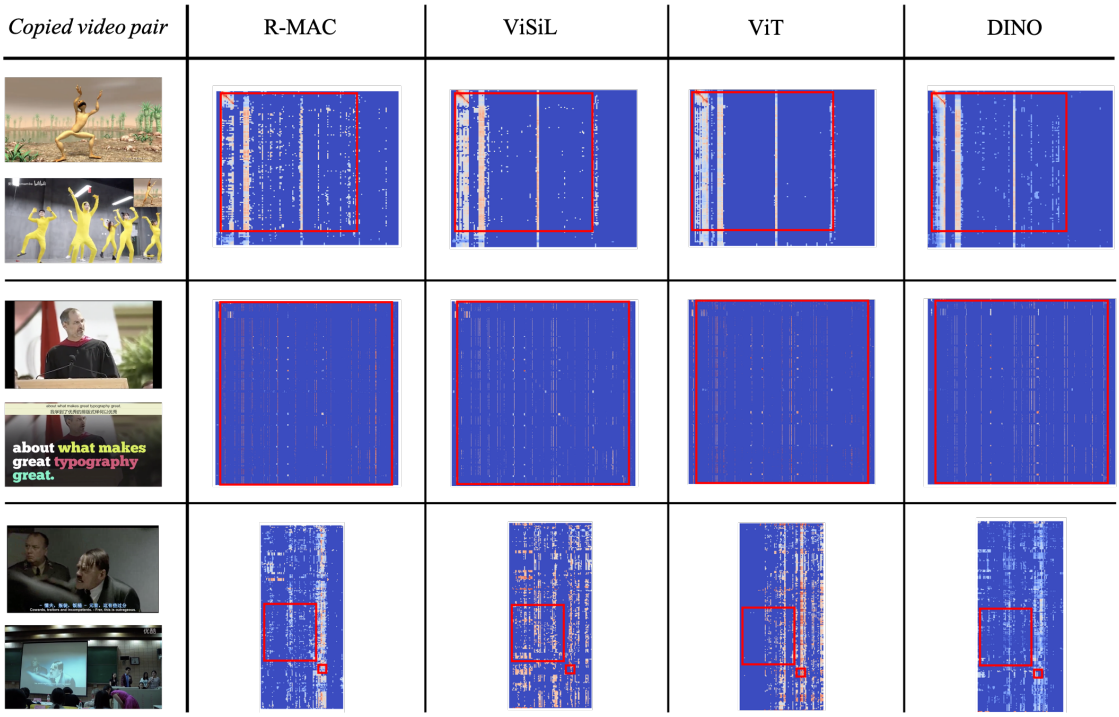

Figure 3. Similarity maps of different features on some hard cases. Red bounding boxes indicate ground-truth copied segment locations. The left column shows the original copied video pairs. Almost all the experimental feature extractors cannot work well on these hard cases.

Table 1. Benchmark F-score results at different topic category

Method	variety show	games	music videos	news	sports	daily life	kichiku	adver-tisement	anima-tion	movies	TV series
HV	81.33	93.34	67.20	77.48	53.57	47.33	**59.96**	67.42	90.67	58.90	91.16
TN	**97.14**	**97.41**	76.88	89.69	59.38	65.52	57.19	75.36	91.43	65.41	**95.68**
DP	81.97	92.02	71.87	83.10	57.10	37.23	46.79	71.13	86.35	60.37	90.26
DTW	76.87	92.75	65.21	72.48	48.98	45.06	53.75	61.30	92.82	48.07	84.67
SPD_1	96.09	94.69	86.52	89.34	58.23	68.37	49.77	**76.39**	93.92	55.10	93.96
SPD_2	96.04	96.49	**89.27**	**92.84**	**70.79**	**71.10**	54.37	74.93	**96.92**	**70.18**	95.46

S4. Benchmark results at different topic category

We show the overall results of all combinations of feature extractors and temporal alignment methods in Section 5.3 of our main text. In this part of Supplementary Material, more fine-grained results are calculated by averaging the metric results of each video pair in different topic categories. The detailed F-score results are shown in Table.1 above with bold text highlighted on the best performance among all the temporal aligned methods. The frame feature extractor here is DINO with better feature representation verified before.

As we mentioned in our main text, results on some specific query sets are only around 50% which are far from satisfactory, especially on some query sets in kichiku and movie category with significant temporal and spatial editing. Among all the temporal alignment algorithms, SPD trained on VCSL achieves the best accuracy results on more than half of the topic categories (6/11), and TN also has three highest scores but only slightly better than SPD_2 on these three categories. It is interesting that the simplest HV method obtains the best result on the most difficult topic kichiku, even though this best result is also less than 60%. Besides the impact of feature extractors, all these temporal alignment methods also need to be specifically optimized for this recently emerged copy infringement types in VCSL.

Table 2. Benchmark F-score results at different video durations

Method	10-30s	30-60s	60s-2min	2-4min	4-8min	8-15min	15-30min	>30min
HV	66.15	64.09	59.30	62.87	59.59	51.75	75.50	59.66
TN	**83.45**	**68.67**	60.90	64.79	70.98	62.45	90.45	69.89
DP	78.04	62.70	53.11	54.51	61.10	52.51	71.58	56.81
DTW	76.11	55.98	51.75	50.33	54.09	41.77	55.61	54.57
SPD_1	75.68	63.01	55.08	60.56	63.33	65.15	88.46	69.67
SPD_2	77.40	67.24	**63.54**	**71.78**	**76.47**	**68.07**	**90.68**	**71.82**

Table 3. Benchmark F-score results at different segment durations

Method	0-5s	5-10s	10-20s	20-45s	45-90s	90s-3min	3-6min	6-30min	>30min
HV	**45.76**	**48.18**	50.12	59.26	65.60	74.59	86.44	79.64	**90.19**
TN	17.50	31.01	49.91	64.81	**71.80**	79.10	88.58	67.44	77.06
DP	19.49	35.39	51.50	56.07	52.95	64.73	83.86	68.98	76.45
DTW	28.61	30.24	40.84	51.96	56.44	69.63	82.84	67.06	76.45
SPD_1	9.99	18.57	38.32	61.30	69.38	83.86	**92.73**	**79.71**	76.09
SPD_2	25.87	43.65	**63.85**	**65.12**	68.47	**85.53**	92.52	77.31	75.84

Table 4. Benchmark F-score results at different segment numbers per video pair

Method	1	2	3	4	5	>5
HV	65.31	52.55	**45.83**	**52.13**	**42.24**	49.25
TN	70.63	51.46	44.75	45.76	37.28	**51.68**
DP	60.91	45.66	36.19	35.26	25.72	36.75
DTW	55.79	41.99	44.79	43.96	41.95	36.49
SPD_1	66.45	41.28	27.76	35.91	19.95	35.90
SPD_2	**74.31**	**56.55**	40.16	42.38	18.93	36.58

Table 5. Benchmark F-score results at different copy duration percentages

Method	0-20%	20-40%	40-60%	60-80%	80-100%
HV	43.67	53.70	66.02	67.99	79.28
TN	43.19	55.86	70.39	**75.06**	84.14
DP	45.09	53.70	57.65	57.16	73.58
DTW	35.19	47.77	49.92	59.83	79.50
SPD_1	22.05	54.87	**71.26**	69.24	87.68
SPD_2	**59.63**	**62.95**	69.19	70.77	**88.30**

S5. Benchmark results at different data distributions

In this section, we evaluate the algorithms at different data distributions corresponding to statistics in Fig.2 of our main text. In detail, F-score performance results on video duration, segment duration, segment number per video pair, copy duration percentage are indicated in Table 2-5 above. Here, the video and segment duration are calculated by averaging the two videos and segments duration of each pair. Similar with Table 1, frame feature extractor here is also DINO.

As can be shown in Table 2, SPD_2 (trained on VCSL) outperforms others on video duration longer than 1 min, and TN performs better on short videos. This is due to the resize operation in the preprocess of SPD, and this significantly rescales the similarity maps of short video pairs. In contrast, TN and other traditional methods without training process directly operate

on the frame similarities, and this is more suitable for short videos. Table 3 shows the benchmark results of different segment durations, and SPD achieves better performance on most of the duration ranges with larger amounts in VCSL (amount data on segment duration can be referred from Fig.2(b) in the main text). For the extreme short segments (<10s) and long segments (>30min), HV method obtains the best result. After manually going through these segments, it is found that entire copied segments covering the diagonal of similarity map occupy most in these extreme cases and HV is suitable for these cases. From Table 4 and Table 5, video pairs containing more segment copies and lower copy duration percentages meet significantly more difficulties with lower results. In Table 5, SPD significantly outperforms other methods on copy duration percentage with larger quantity of video pair data (0-20% and 80-100%, indicated from Fig.2(d) in the main text). This might also be attributed to the scale and diversity of training dataset of VCSL. Overall, large-scale and well-annotated datasets are essential for supervised learning methods, and temporal alignment methods show different adaptability on various data distributions and situations.

S6. Limitation of metric on extreme cases

We propose our new metric in Section.4 of the main text, and this metric fully considers the segment division equivalence of copy detection task. But this consideration also brings limitations on some rare and extreme cases shown in Fig.4 below. As we mentioned after Eq.(4) of main text, we utilize the projection length on x and y axis rather than the bounding box area that is more commonly used in IoU. This is to make the metric more robust against the equivalence of a single bounding box and its division of temporally consecutive bounding boxes shown in Fig.4(a). However, the wrong prediction with reversed order can also be measured with high scores shown in Fig.4(b), and extra prediction boxes inside ground truth box are also not punished by our metric shown in Fig.4(c). This is the trade-off between robustness on segment division and discrimination on inner inaccurate prediction. This trade-off cannot be perfectly solved since even the most precise annotation can only be finished at segment-level (boxes in similarity map) rather than frame-level (points in the similarity map) while considering annotation cost and operability. After observation on the prediction results, we find that the cases similar with Fig.4(a) appear far more frequently than Fig.4(b) and Fig.4(c). Similarity maps with a large GT box containing several inaccurate predicted boxes rarely happen with an optimized temporal alignment method. Therefore, we have to sacrifice the measurement accuracy on the rare cases like Fig.4(b) and (c), and make our metric to be robust against common segment division shown in Fig.4(a).

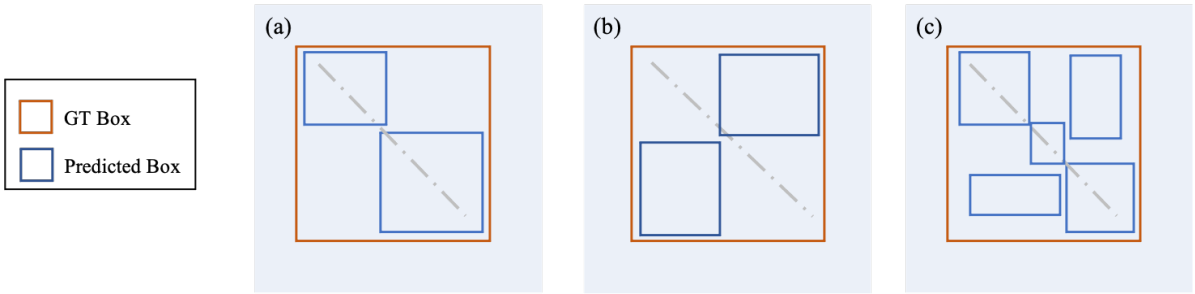

Figure 4. All the above situations are evaluated with high precision and recall results by our metric. However, (b) and (c) predict inaccurate localization. In detail, (a)correct measurement of our metric considering segment division equivalence (same with Fig.4(d) in main text); (b)inaccurate measurement with reversed-order predicted boxes; (c)inaccurate measurement with extra wrong predicted boxes.

References

[1] Giorgos Kordopatis-Zilos, Symeon Papadopoulos, Ioannis Patras, and Ioannis Kompatsiaris. Visil: Fine-grained spatio-temporal video similarity learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6351–6360, 2019. 1

[2] Chen Jiang, Kaiming Huang, Sifeng He, Xudong Yang, Wei Zhang, Xiaobo Zhang, Yuan Cheng, Lei Yang, Qing Wang, Furong Xu, et al. Learning segment similarity and alignment in large-scale content based video retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1618–1626, 2021. 1