A. Details about Eq. 7

A.1. Proof of Eq. 7 in the main paper

$$\sigma\left(v_i|i\in[1,n]\right)=\exp\left(\sum_{i=1}^{n}\sum_{j<i}^{n}\left(v_i^Tv_j\right)\right) \tag{1}$$
$$=exp\left(\left\|\alpha\right\|^2/2\right)\cdot exp\left(-\left\|v_1\right\|^2/2\right)\cdot\ldots\cdot exp\left(-\left\|v_n\right\|^2/2\right)$$

where $\alpha=(v_1+\ldots+v_n)$. Next, let $\omega\in\mathbb{R}^{K^{'}}$. We use the fact that:

$$(2\pi)^{-K^{'}/2}\int exp\left(-\left\|\omega-c\right\|_2^2/2\right)d\omega=1 \tag{2}$$

For any $c\in\mathbb{R}^{K^{'}}$ and derive:

$$exp\left(\left\|\alpha\right\|^2/2\right)$$
$$=(2\pi)^{-K^{'}/2}exp\left(\left\|\alpha\right\|^2/2\right)\int exp\left(-\left\|\omega-\alpha\right\|^2/2\right)d\omega$$
$$=(2\pi)^{-K^{'}/2}\int exp\left(-\left\|\omega\right\|^2/2+\omega^T\alpha-\left\|\alpha\right\|^2/2+\left\|\alpha\right\|^2/2\right)d\omega$$
$$=(2\pi)^{-K^{'}/2}\int exp\left(-\left\|\omega\right\|^2/2+\omega^T\alpha\right)d\omega$$
$$=(2\pi)^{-K^{'}/2}\int exp\left(-\left\|\omega\right\|^2/2\right)\cdot exp\left(\omega^Tv_1\right)\cdot\ldots\cdot exp\left(\omega^Tv_n\right)d\omega$$
$$=\mathbb{E}_{\omega\sim\mathcal{N}(0,\mathbf{I_{K'}})}\left[exp\left(\omega^Tv_1\right)\cdot\ldots\cdot exp\left(\omega^Tv_n\right)\right] \tag{3}$$

That completes the proof.

A.2. Theoretical Error of Eq. 7 in the main paper

$$\sigma\left(v_i|i\in[1,n]\right)=exp\left(-\frac{\beta}{2}\right)\mathbb{E}_{\omega\sim\mathcal{N}(0,\mathbf{I_{K'}})}\left[exp\left(\omega^T\alpha\right)\right] \tag{4}$$

where $\beta=\left(\left\|v_1\right\|^2+\ldots+\left\|v_n\right\|^2\right)$, based on the fact $\mathbb{E}_{\omega\sim\mathcal{N}(0,\mathbf{I_{K'}})}\left[exp\left(\omega^T\alpha\right)\right]=exp\left(\frac{\left\|\alpha\right\|^2}{2}\right)$, then we can obtain:

$$\mathrm{MSE}\left(\sigma\left(v_i|i\in[1,n]\right)\right)=\frac{1}{H}exp\left(-\beta\right)\mathrm{Var}\left(exp\left(\omega^T\alpha\right)\right)$$
$$=\frac{1}{H}exp\left(-\beta\right)\left(\mathbb{E}\left[exp\left(2\omega^T\alpha\right)\right]-\left(\mathbb{E}\left[exp\left(\omega^T\alpha\right)\right]\right)^2\right)$$
$$=\frac{1}{H}exp\left(-\beta\right)\left(exp\left(2\left\|\alpha\right\|^2\right)-exp\left(\alpha^2\right)\right) \tag{5}$$
$$=\frac{1}{H}exp\left(-\beta\right)exp\left(\left\|\alpha\right\|^2\right)\left(exp\left(\left\|\alpha\right\|^2\right)-1\right)$$
$$=\frac{1}{H}exp\left(\left\|\alpha\right\|^2\right)\sigma^2\left(v_i|i\in[1,n]\right)\left(1-exp\left(-\left\|\alpha\right\|^2\right)\right)$$

where H denotes the number of random features.

B. Details about Average

In this study, we focus on identifying which frames correspond to the shared content beforehand by considering the relationships among videos in this study. The intuitive approach would be to summarize the target and reference features by averaging, and concatenating them to create the final feature for description generation. The process can be formalized as follows.

we first average the video features in the time direction as below:

$$v_i^{'}=\frac{1}{m}\sum_{j=1}^{m}v_i^j \tag{6}$$

Then we simply average the feature $v_i^{'}$ over n videos because the shared content should appear in all videos in an input group.

$$\psi_{tar}=\frac{1}{n_{tar}}\sum_{i=1}^{n_{tar}}v_i^{'}\qquad\psi_{ref}=\frac{1}{n_{ref}}\sum_{i=1}^{n_{ref}}v_i^{'} \tag{7}$$

Finally, we concatenate ψ_{tar} and ψ_{ref} and input it to the decoder as described in Eq. 4.

C. Comparison with Single Video Captioning

In this section, we describe the difference between our group captioning task and the existing individual image captioning task. Captioning the group as a whole is different from processing each video individually and then summarizing them can not solve our task. We use the state-of-the-art video captioning model, HMN[1], to generate the individual video captions.

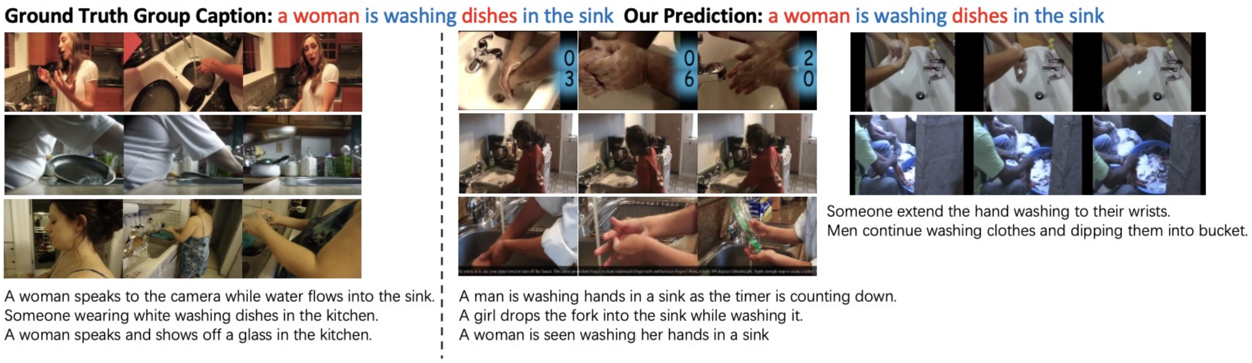
In order to quantitatively compare group-based video captioning methods with existing single video captioning methods, we compare the generated individual descriptions by HMN with ground-truth sentences and used the mean evaluation score of the target group as the final score as shown in Table 1. The result (Per-Video) indicates that the group captioning problem cannot be solved by simply summarizing per-image captions.

Fig 1 shows one example from the ActivityNet Caption dataset. The 3 captions on the right correspond to the target videos on the left in order, and the 5 captions on the right correspond to the reference videos on the right in order. While the video group features for **dishes** and **in the sink**, individual captions focus on other aspects including with **water flows**, **in the kitchen** or **shows off glass**. Only one per-video caption notices that the woman is washing dishes. Therefore, if we are summarizing the target per-video captions to get group caption, we will get result **A**

[1] https://github.com/MarcusNerva/HMN

Dataset	Methods	B@1	B@2	B@3	B@4	WAC	METEOR	ROUGE	CIDEr	WER↓
YouCook2	Per-Video	16.99	9.31	4.27	1.64	12.71	6.86	13.29	53.67	140.75
	Average	44.93	28.77	19.03	12.69	34.13	20.72	43.86	170.6	85.34
	Traversal	46.15	29.52	**21.22**	14.04	38.80	21.84	**45.97**	179.4	**75.61**
	ERA	**47.32**	**30.74**	20.78	**14.74**	**39.55**	**22.18**	45.77	**180.8**	75.77
ActivityNet	Per-Video	15.41	10.12	3.75	1.86	12.99	7.13	14.32	55.27	138.48
	Average	41.87	28.46	19.53	16.17	33.62	20.32	40.17	172.8	86.01
	Traversal	**44.57**	29.70	21.28	16.53	37.85	**21.64**	**42.78**	180.7	74.94
	ERA	44.26	**29.75**	**21.61**	**16.97**	**37.93**	21.51	42.42	**181.3**	**74.41**

Table 1. Comparison with single video captioning


Ground Truth Group Caption: a woman is washing dishes in the sink **Our Prediction:** a woman is washing dishes in the sink

A woman speaks to the camera while water flows into the sink.
Someone wearing white washing dishes in the kitchen.
A woman speaks and shows off a glass in the kitchen.

A man is washing hands in a sink as the timer is counting down.
A girl drops the fork into the sink while washing it.
A woman is seen washing her hands in a sink

Someone extend the hand washing to their wrists.
Men continue washing clothes and dipping them into bucket.

Figure 1. Individual captions generated by single video captioning model, HMN.

woman in the kitchen, which misses out the most important feature of the video group(washing dishes in the sink).

The information needed for group captioning may be missed out in individual captions because the common feature of the group might not be important for individual videos. Therefore, captioning the group as a whole is different from processing each image individually and then summarizing them. This also explains why end-to-end captioning generation for a video group can capture information that individual captions tend to miss and resulting in more informative group captions.

D. Experiments

D.1. Varying the Number of Reference Images

In Table 5 of the main paper, we give experimental results of varying the number of target and reference videos on YouCook2. Here in Table 2 we give more results by varying the number of videos on YouCook2. We also give the results evaluated on the ActivityNet Caption dataset as shown in Table 3. As shown in the tables, the performance improves when more reference videos are given. We also notice that while the differences between giving 0, 1, or 3 reference videos are large, the gap between 3 and 5 is insignificant. So we use 5 reference videos in the overall experiment setting.

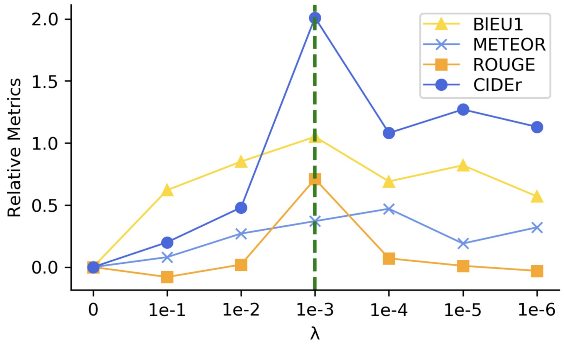


Figure 2. Model performances under different λ hyper-parameters on the YouCook2 dataset. The best model performances are marked in the green line.

D.2. Hyper-Parameter Analysis

To evaluate the impact of λ and find an appropriate ratio $\mathcal{L}_{cap}$ and $\mathcal{L}_{igc}$, we adjust the value of λ in Equation 12 based on ERA. The results are shown in Figure 2 and Figure 3. We observe that an intra-contrastive refinery with $\lambda = 1e - 3$ achieves the best performance. In addition, removing intra-contrastive ($\lambda = 0$) shows the worst performance, indicating that our proposed strategy is a significant boost for integrating the OCR words correctly.

Methods	BLEU@1	BLEU@2	BLEU@3	BLEU@4	WAC	METEOR	ROUGE	CIDEr	WER↓
T1+R5	41.94	26.56	16.47	14.83	31.63	19.86	40.86	163.1	80.43
T2+R5	45.64	27.76	17.04	14.54	37.36	21.68	43.84	174.3	77.23
T3+R0	19.87	12.52	9.88	8.32	13.92	8.63	19.15	69.81	97.62
T3+R1	38.45	19.21	14.65	11.02	22.52	15.23	33.47	112.6	93.58
T3+R3	40.09	23.34	17.27	13.19	27.53	17.58	39.83	158.5	89.47
T3+R5	**47.32**	**30.74**	**20.78**	**14.74**	**39.55**	**22.18**	**45.77**	**180.8**	**75.77**

Table 2. Performance with varying the number of target and reference videos. (evaluated on YouCook2 Captions dataset)

Methods	BLEU@1	BLEU@2	BLEU@3	BLEU@4	WAC	METEOR	ROUGE	CIDEr	WER↓
T1+R5	40.31	26.81	19.43	14.66	34.23	20.61	39.64	164.7	78.41
T2+R5	42.06	28.21	20.74	15.09	36.63	21.11	41.32	176.6	76.18
T3+R0	21.52	13.64	10.68	9.41	15.22	10.53	24.99	72.74	96.28
T3+R1	31.22	19.09	15.19	12.56	24.56.	15.05	31.47	115.6	89.84
T3+R3	38.73	24.74	18.59	15.01	28.34	18.98	37.64	157.4	80.34
T3+R5	**44.26**	**29.75**	**21.61**	**16.97**	**37.93**	**21.51**	**42.42**	**181.3**	**74.41**

Table 3. Performance with varying the number of target and reference videos. (evaluated on ActivityNet Caption dataset)

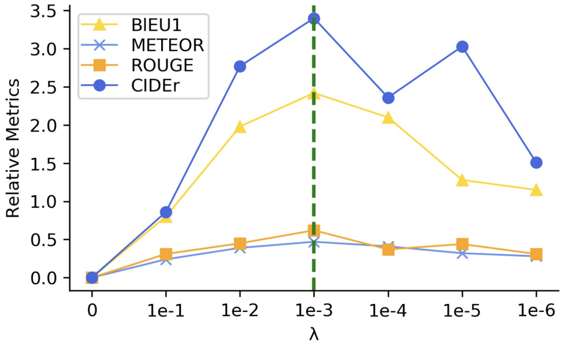


Figure 3. Model performances under different λ hyper-parameters on ActivityNet Caption dataset. The best model performances are marked in the green line.

E. Application of Group Video Captioning

Application scenarios discussion. In real-word, there are many promising application scenarios for group video captioning. For example, applications include **1)** titling categorized video folder for social sharing **2)** query suggestions for text-based video retrieval **3)** video recommendation reason generation based on user's historical behavior. **4)** In addition, we hope the proposed datasets and ERA module can also be extended to group-based video analysis tasks like group-based video QA, and bring more inspiration to the community.

Task setting discussion. We set the same number of videos for each group (3 for target videos and 5 for reference videos) just for experimental purposes, in practice, this can be adjusted as needed. And, there are many ways to meet the condition. When there are too many videos, the videos can be sorted according to different strategies, and when the videos are not enough, the videos can be expanded by data enhancement like flipping, scaling, or adjusting the frame sampling interval.

F. Limitation

In this paper, we establish the first group video captioning benchmark. We aim to describe a group of target videos in the context of another group of related reference videos and construct two group video captioning dataset. The proposed method provides a possible new approach to generating precise and relevant sentences for video groups and may inspire more work. It may also help to develop more practical video processing systems. However, this technique still suffers from biases in the training data. It may produce incorrect output or lead to an inaccurate understanding of video content when the video involves uncommonly-seen subjects. Therefore, more research is necessary to address this issue in the future.

G. More Examples

Figure 4 and Figure 5 show more good examples on YouCook2 and ActivityNet Caption datasets respectively. Figure 6 and Figure 7 show failure cases on the two datasets. Analysis for the failure cases (Figure 6, Figure 7) can be found in the captions of each figure.

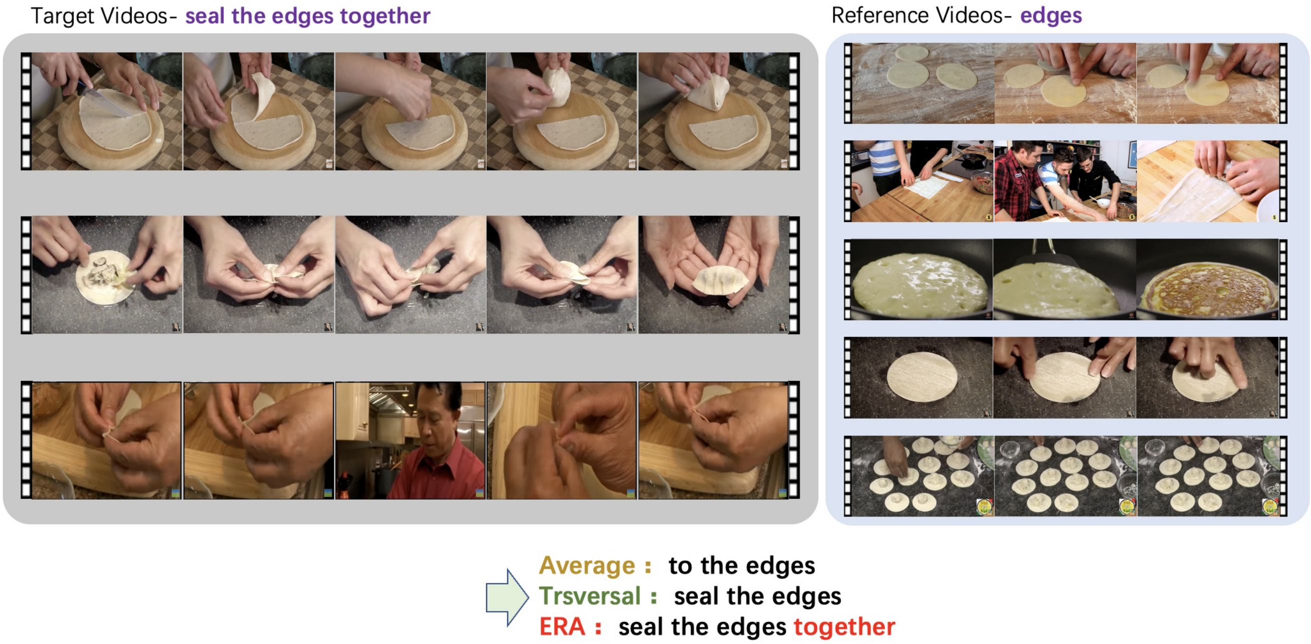

Target Videos- seal the edges together
Reference Videos- edges
Average : to the edges
Trsversal : seal the edges
ERA : seal the edges together

Figure 4. Good examples on YouCook2 dataset.

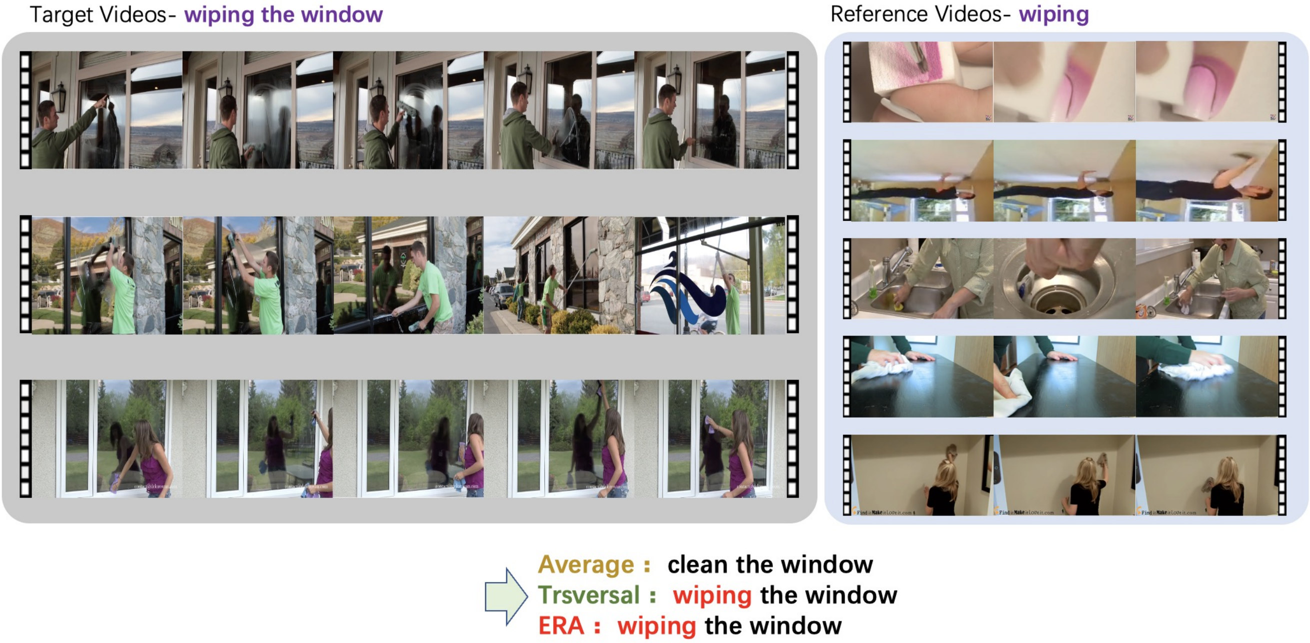

Target Videos- wiping the window
Reference Videos- wiping
Average : clean the window
Trsversal : wiping the window
ERA : wiping the window

Figure 5. Good examples on ActivityNet Caption dataset.

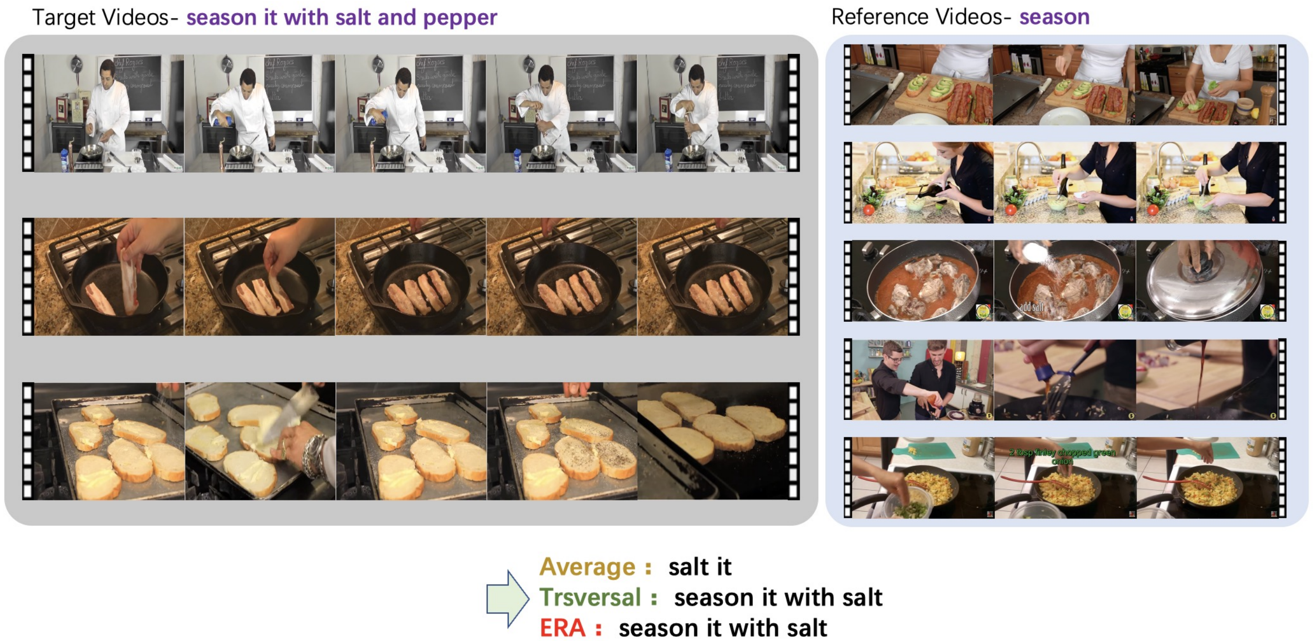

Target Videos- season it with salt and pepper
Reference Videos- season
Average : salt it
Trsversal : season it with salt
ERA : season it with salt

Figure 6. Failure cases on the YouCook2 dataset. The model only predicts **salt** missing **pepper** which is visually similar. This may be because the model does not capture features of the **pepper** well.

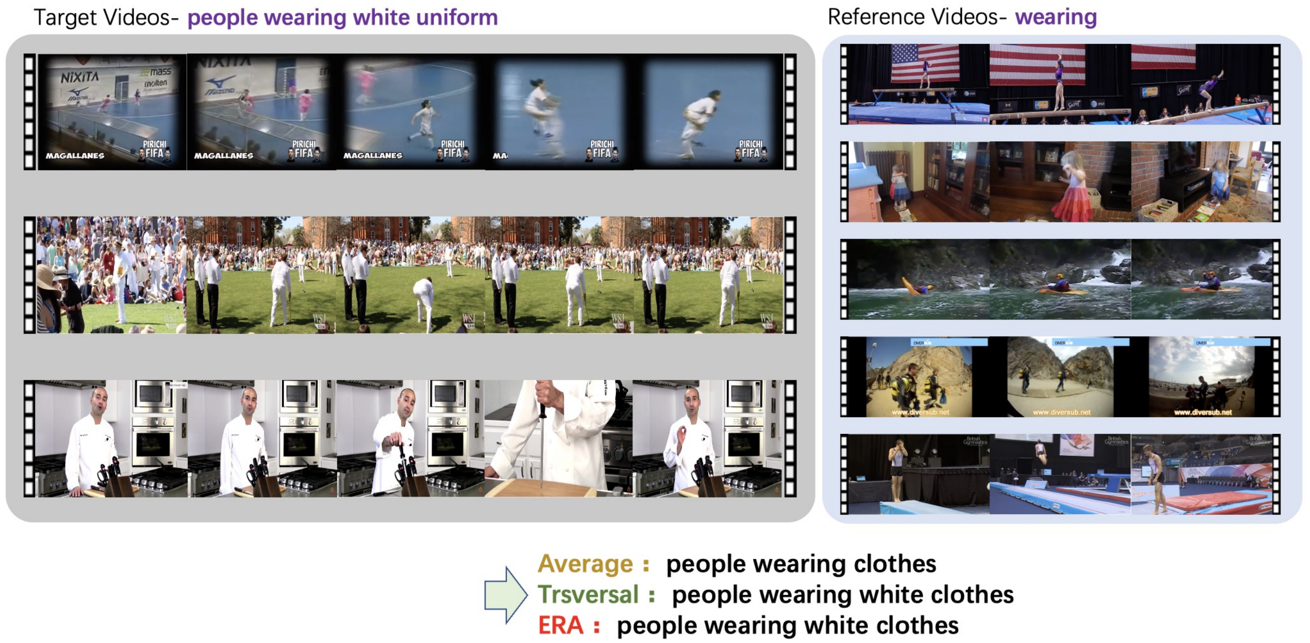

Target Videos- people wearing white uniform
Reference Videos- wearing
Average : people wearing clothes
Trsversal : people wearing white clothes
ERA : people wearing white clothes

Figure 7. Failure cases on ActivityNet Caption dataset. For this example, the model prediction **clothes** is correct but not as good as ground-truth **uniform**.