Supplementary Material for
Hierarchical Video-Moment Retrieval and Step-Captioning

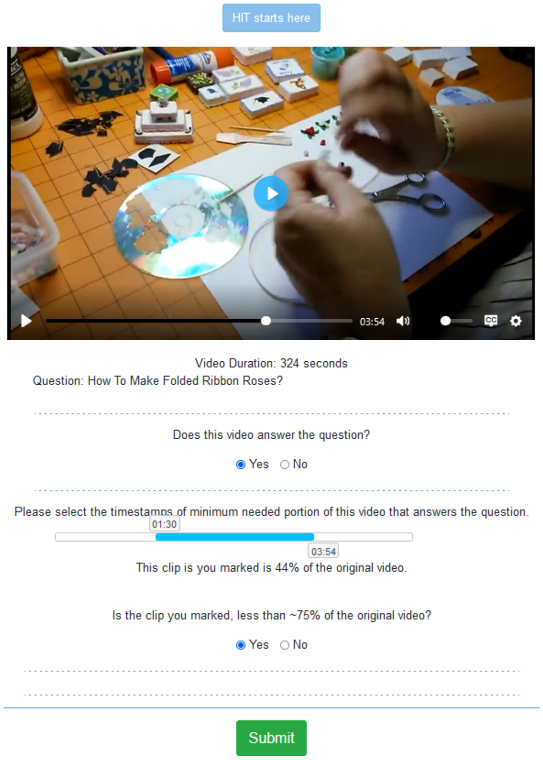
Abhay Zala* [1] Jaemin Cho* [1] Satwik Kottur[2] Xilun Chen[2]
Barlas Oguz[2] Yashar Mehdad[2] Mohit Bansal[1]
UNC Chapel Hill[1] Meta AI[2]
{jmincho, aszala, mbansal}@cs.unc.edu {skottur, xilun, barlaso, mehdad}@fb.com

In the appendix, we include the following content: Data annotation details (Appendix A), CLIP-based moment retrieval method visualization (Appendix B), Model performance analysis in video categories and duration groups (Appendix C), and Evaluation details (Appendix D).

A. Data Annotation Details

Annotation Interface. In the following, we provide screenshots of the HIREST annotation interface for each stage (Fig. 7 and Fig. 8) and worker qualification process.

Worker Qualifications and Pay. We require crowdworkers to have above a 95% approval rate and have completed at least 1000 or more other tasks before working on ours. We also require that all workers pass a qualification test (separate tests for each stage) before they can work on our tasks. For stage 1, the qualification test is composed of 2 parts. First workers are asked to determine if a video solves/answers the given prompt, then they are shown a relevant video and asked to identify the relevant moment in the video (we provide some leniency with timing). For stage 2, workers are given a short series of videos with multiple-choice questions. The multiple choice answers consist of pre-written step captions for the video. The workers were asked then to identify which set of step captions was the best (*i.e*. best covered the video and didn't violate simple rules). A total of 72 workers passed the qualification test for both tasks. As text queries from the HowTo100M [4] dataset are all in English and all of our collected step captions are in English, we require crowdworkers to be from an English-speaking country. Workers were paid $0.20 for stage 1 and $0.45 for stage 2. We also provide a large bonus for good workers. For stage 1, workers are bonused with $0.05 if the video answers/solves the prompt and if they correctly trim the video. Then for every 25 tasks they complete, their base pay increases by $0.02. A typical worker can earn up to $0.27 per task, which is roughly $12.00 per hour. For stage 2, workers are paid with $0.04 bonus for every high-quality step caption they write complete and for every 10 high-quality tasks they complete, their base pay is increased by $0.02. A typical worker can earn up to $0.81 per task, which is roughly $12.00 per hour. For both tasks, there is a baseline pay of around $12, but oftentimes, workers would complete more than 25 and 10 tasks (respectively, for stages 1 and 2), pushing the pay/hour higher.



Video Duration: 324 seconds
Question: How To Make Folded Ribbon Roses?

Does this video answer the question?

● Yes ○ No

Please select the timestamps of minimum needed portion of this video that answers the question.

01:30 03:54

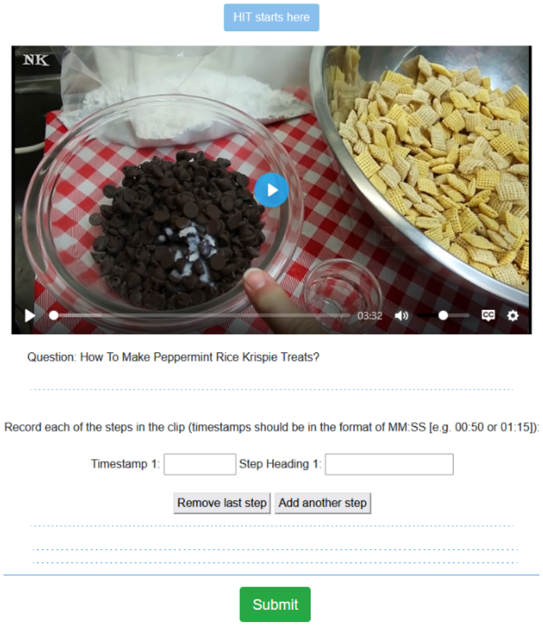
This clip is you marked is 44% of the original video.

Is the clip you marked, less than ~75% of the original video?

● Yes ○ No

Figure 7. Stage 1 data collection interface for video and moment retrieval. Crowdworkers are presented with a video and text query and asked if the video answers/solves the question. If they select yes, the interface expands to a sliding bar that allows them to trim the video down to just the relevant portion.

[*] equal contribution



Question: How To Make Peppermint Rice Krispie Treats?

Record each of the steps in the clip (timestamps should be in the format of MM:SS [e.g. 00:50 or 01:15]):

Timestamp 1: Step Heading 1:

Figure 8. Stage 2 data collection interface moment segmentation and step captions. Crowdworkers are presented with a video and instructional text query and are asked to write all the essential steps in the video along with the timestamp of each step.

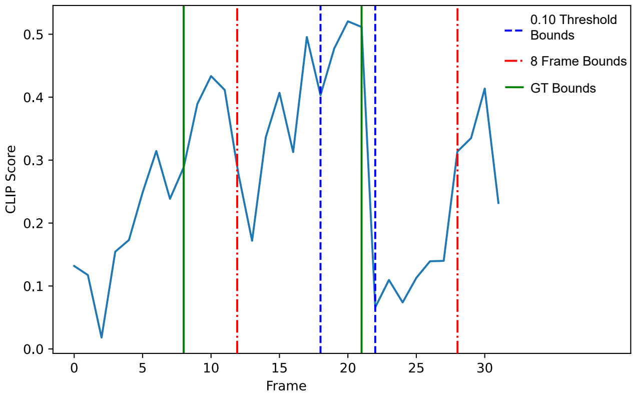


Figure 9. Visualization of image-text cosine similarity-based methods for the moment retrieval task. The 8-frame method (red with dashes and dots) achieves IoU=0.42 while the 0.10 threshold method (blue with dashes) achieves IoU=0.19. Ground Truth bounds are indicated with the solid green lines. EVA-CLIP model was used for the plot. *CLIP Score: cosine similarity between image and text embedding*.

Category	# Prompts	# Videos
Hobbies and Crafts	193	231
Food and Entertaining	192	250
Home and Garden	69	111
Cars and Other Vehicles	28	55
Holidays and Traditions	25	47
Education and Communications	15	23
Personal Care and Style	6	29
Pets and Animals	5	6
Health	5	13
Family Life	4	1
Arts and Entertainment	1	1
Sports and Fitness	1	8
Misc.	2	1
All	546	776

Table 7. Prompt and Video category distributions of HIREST test split. Categories are sorted in descending order by the number of prompts. The number of prompts is smaller than the number of videos since multiple videos were retrieved and paired with some prompts.

Category	Model	FT	R@1	R@5	R@10
Hobbies and Crafts	EVA-CLIP		26.42	52.85	63.73
Food and Entertaining	EVA-CLIP		25.52	43.75	53.12
Home and Garden	EVA-CLIP		27.54	62.32	71.01
Cars and Other Vehicles	EVA-CLIP		14.29	53.57	64.29
Holidays and Traditions	EVA-CLIP		44.0	68.0	76.0
Education and Communications	EVA-CLIP		20.0	26.67	46.67
Personal Care and Style	EVA-CLIP		50.0	66.67	66.67
Pets and Animals	EVA-CLIP		0	60.0	80.0
Health	EVA-CLIP		60.0	60.0	80.0
Family Life	EVA-CLIP		0	75.0	100.0
Arts and Entertainment	EVA-CLIP		0	0	0
Sports and Fitness	EVA-CLIP		100.0	100.0	100.0
All	EVA-CLIP		26.37	51.1	61.54

Table 8. Video retrieval results per prompt category on our HIREST test split. *FT: finetuning on* HIREST, *R@k: Recall@k.*

B. CLIP-based Moment Retrieval Method

In Fig. 9, we illustrate two heuristics that we discuss in the main paper Sec. 4.1. From the frame that scores the highest text-frame cosine similarity, we determine the start/end timestamp of moment by 1) picking the frames where the similarity score drops from the highest scoring frame by a certain threshold (*e.g.*, 0.10); 2) picking the 8 frames to the left and right, totaling up to 17 (= 8+1+8) frames.

C. Model Performance Analysis in Video Categories and Duration

As mentioned in the main paper Sec. 3, HIREST videos are collected by text queries with different categories such as ‘Home and Garden’ and ‘Food and Entertaining’. Also, videos and moments have different durations. In Table 7, we show the distribution of prompts and videos for each category in HIREST test split. In the following, we provide comprehensive evaluation results per category and different video/moment duration groups.

Category	Model	FT	R@0.5	R@0.7
Hobbies and Crafts	BMT		50.65	11.26
Food and Entertaining	BMT		34.40	8.80
Home and Garden	BMT		39.64	6.31
Cars and Other Vehicles	BMT		60.00	14.55
Holidays and Traditions	BMT		36.17	6.38
Education and Communications	BMT		47.83	26.09
Personal Care and Style	BMT		44.83	10.34
Pets and Animals	BMT		33.33	33.33
Health	BMT		61.54	30.77
Family Life	BMT		0	0
Arts and Entertainment	BMT		100	0
Sports and Fitness	BMT		62.5	12.5
All	BMT		43.56	10.57
Hobbies and Crafts	BMT	✓	72.29	39.39
Food and Entertaining	BMT	✓	72.80	38.00
Home and Garden	BMT	✓	67.57	36.04
Cars and Other Vehicles	BMT	✓	74.55	52.72
Holidays and Traditions	BMT	✓	72.34	31.91
Education and Communications	BMT	✓	60.87	39.13
Personal Care and Style	BMT	✓	72.41	34.38
Pets and Animals	BMT	✓	66.67	16.67
Health	BMT	✓	84.62	69.23
Family Life	BMT	✓	100	100
Arts and Entertainment	BMT	✓	100	100
Sports and Fitness	BMT	✓	87.5	37.5
Hobbies and Crafts	Joint (Ours)	✓	75.76	35.5
Food and Entertaining	Joint (Ours)	✓	75.2	36.4
Home and Garden	Joint (Ours)	✓	63.06	21.62
Cars and Other Vehicles	Joint (Ours)	✓	81.82	34.55
Holidays and Traditions	Joint (Ours)	✓	72.34	31.91
Education and Communications	Joint (Ours)	✓	78.26	26.09
Personal Care and Style	Joint (Ours)	✓	68.97	17.24
Pets and Animals	Joint (Ours)	✓	33.33	33.33
Health	Joint (Ours)	✓	61.54	30.77
Family Life	Joint (Ours)	✓	100.0	100.0
Arts and Entertainment	Joint (Ours)	✓	100.0	0.0
Sports and Fitness	Joint (Ours)	✓	75.0	50.0
All	BMT	✓	71.91	39.18
All	Joint (Ours)	✓	73.32	32.60

Table 9. Moment retrieval results per video category on our HiREST test split. *FT: Finetuning on* HiREST*, R@IoU: Recall@1 with a threshold of IoU.*

Video Duration	Model	FT	R@0.5	R@0.7
$<$ 2 mins	BMT		48.70	10.43
2 - 6 mins	BMT		37.47	7.04
$>$ 6 mins	BMT		56.74	20.22
All	BMT		43.56	10.57
$<$ 2 mins	BMT	✓	74.78	44.35
2 - 6 mins	BMT	✓	72.05	40.37
$>$ 6 mins	BMT	✓	69.66	32.58
$<$ 2 mins	Joint (Ours)	✓	68.10	18.10
2 - 6 mins	Joint (Ours)	✓	73.21	28.21
$>$ 6 mins	Joint (Ours)	✓	75.00	40.26
All	BMT	✓	71.91	39.18
All	Joint (Ours)	✓	73.32	32.60

Table 10. Moment retrieval results for various durations on our HiREST test split. *FT: Finetuning on* HiREST*, R@IoU: Recall@1 with a threshold of IoU.*

Video retrieval. In Table 8, we show EVA-CLIP [1] (ViT-G/14) with 20 frames on each prompt category in our dataset. Among the categories that have many most videos ($>$ 20 videos), the model is better at 'Holidays and Traditions' and 'Personal Care and Style' than 'Cars and Other Vehicles'.

Moment retrieval. In Table 9, we show the zeroshot and finetuning results of BMT [2] proposal module and our joint model on each HiREST video category. Categories like 'Home and Garden' and 'Holidays and Traditions' see a strong performance increase after finetuning.

In Table 10, we show Moment retrieval performance on three video duration groups. Before finetuning, BMT performs slightly better on videos of a longer length; however, after finetuning, BMT performs much better on shorter-length videos. In R@0.5, our joint model outperforms BMT when the videos are longer.

Moment segmentation. In Table 11, we show the zeroshot and finetuned results of BMT [2] proposal module and our joint model on each individual category in our dataset. Finetuning BMT results show significant improvement in every category.

In Table 12, we show the moment segmentation performance on three moment duration groups. All models achieve higher performance in shorter moments than in longer moments. Our joint model shows better performance than BMT on shorter moments, while BMT does better on longer moments.

Step captioning. In Table 13, we show the zeroshot and finetuned results of SwinBERT [3] and our joint model on each video category in our dataset. Notably, SwinBERT performs best in the 'Food and Entertaining' category, likely because SwinBERT was pretrained on the YouCook2 [5] dataset.

In Table 14, we show the step captioning performance in different step durations. Both N-gram (*e.g.* CIDEr) and sentence-level embedding metrics (BERTScore and CLIPScore) do not show significant differences among different categories. In the entailment metric, finetuned SwinBERT gets better as the steps get longer, while our joint model gets slightly worse for longer steps.

D. Evaluation Details

We continue the evaluation details of moment segmentation task in the main paper Sec. 4.3. Metrics. The BMT [2]

Category	Model	FT	Recall@IoU		Precision@IoU	
			0.5	0.7	0.5	0.7
Hobbies and Crafts	BMT		8.91	3.02	22.44	6.17
Food and Entertaining	BMT		7.47	4.47	19.06	9.36
Home and Garden	BMT		8.64	3.18	23.33	7.62
Cars and Other Vehicles	BMT		4.18	1.11	16.16	5.05
Holidays and Traditions	BMT		7.24	4.67	16.11	10.50
Education and Communications	BMT		6.11	0	20.00	0
Personal Care and Style	BMT		9.87	6.75	19.65	11.58
Pets and Animals	BMT		40.00	10.00	80.00	20.00
Health	BMT		13.33	10.00	30.00	20.00
Family Life	BMT		0	0	0	0
Arts and Entertainment	BMT		0	0	0	0
Sports and Fitness	BMT		0	0	0	0
All	BMT		8.24	3.71	20.95	7.96
Hobbies and Crafts	BMT	✓	33.09	10.13	25.35	7.84
Food and Entertaining	BMT	✓	32.25	12.02	21.86	7.66
Home and Garden	BMT	✓	38.21	13.89	29.04	11.86
Cars and Other Vehicles	BMT	✓	33.09	14.79	22.03	10.60
Holidays and Traditions	BMT	✓	34.57	12.49	26.72	10.00
Education and Communications	BMT	✓	41.47	13.00	23.95	7.07
Personal Care and Style	BMT	✓	29.79	11.47	22.25	7.32
Pets and Animals	BMT	✓	50.00	27.50	42.04	16.86
Health	BMT	✓	42.52	22.30	31.11	18.34
Family Life	BMT	✓	35.71	14.29	38.46	15.38
Arts and Entertainment	BMT	✓	22.22	11.11	11.76	5.88
Sports and Fitness	BMT	✓	33.62	20.13	29.05	11.59
Hobbies and Crafts	Joint (Ours)	✓	38.11	13.63	26.47	9.34
Food and Entertaining	Joint (Ours)	✓	35.43	13.56	28.54	10.4
Home and Garden	Joint (Ours)	✓	35.28	17.63	29.69	14.46
Cars and Other Vehicles	Joint (Ours)	✓	39.68	15.23	31.16	12.39
Holidays and Traditions	Joint (Ours)	✓	34.52	10.92	25.42	6.95
Education and Communications	Joint (Ours)	✓	49.71	24.86	31.61	13.94
Personal Care and Style	Joint (Ours)	✓	41.38	17.57	29.91	13.34
Pets and Animals	Joint (Ours)	✓	70.0	30.0	40.76	17.33
Health	Joint (Ours)	✓	40.58	11.48	37.26	7.42
Family Life	Joint (Ours)	✓	50.0	28.57	63.64	36.36
Arts and Entertainment	Joint (Ours)	✓	22.22	11.11	13.33	6.67
Sports and Fitness	Joint (Ours)	✓	31.71	16.85	24.84	9.5
All	BMT	✓	34.06	12.34	24.71	8.93
All	Joint (Ours)	✓	37.50	14.76	28.52	10.84

Table 11. Moment segmentation results per video category on our HiREST test split. *FT: Finetuning on* HiREST, *Recall@IoU: Recall@1 with a threshold of IoU, Precision@IoU: Precision@1 with a threshold of IoU.*

Moment Duration	Model	FT	Recall@IoU		Precision@IoU	
			0.5	0.7	0.5	0.7
< 1.5 mins	BMT		11.75	4.76	26.92	9.19
1.5 - 3 mins	BMT		6.59	3.31	17.13	7.18
> 3 mins	BMT		6.40	3.04	19.28	7.59
All	BMT		8.24	3.71	20.95	7.96
< 1.5 mins	BMT	✓	38.08	14.27	27.75	10.87
1.5 - 3 mins	BMT	✓	33.92	13.26	23.20	8.74
> 3 mins	BMT	✓	29.54	8.82	23.23	6.90
< 1.5 mins	Joint (Ours)	✓	44.32	17.81	42.22	16.39
1.5 - 3 mins	Joint (Ours)	✓	38.04	14.70	25.41	9.68
> 3 mins	Joint (Ours)	✓	28.72	11.27	16.75	5.93
All	BMT	✓	34.06	12.34	24.71	8.93
All	Joint (Ours)	✓	37.50	14.76	28.52	10.84

Table 12. Moment segmentation results for different moment duration groups on our HiREST test split. *FT: Finetuning on* HiREST, *Recall@IoU: Recall@1 with a threshold of IoU, Precision@IoU: Precision@1 with a threshold of IoU.*

model generates up to 100 possible step segments, where many of them are outside of the ground-truth (GT) moment input, overlap each other, and there are also gaps between segments. For evaluation of moment segmentation task, we first remove any segments outside the given ground truth moment and use non-maximum suppression (NMS) to remove any overlapping segments. Then any resulting gaps between steps are also marked as separate steps.

References

[1] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *CVPR*, 2023. 3

[2] Vladimir Iashin and Esa Rahtu. A better use of audio-visual cues: Dense video captioning with bi-modal transformer. In *British Machine Vision Conference (BMVC)*, 2020. 3

[3] Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. Swinbert: End-to-end transformers with sparse attention for video captioning. In *CVPR*, 2022. 3

[4] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*, 2019. 1

[5] Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018. 3

Category	Model	FT	METEOR	CIDEr	SPICE	Entailment (%)	BERTScore	CLIPScore
Hobbies and Crafts	SwinBERT		3.18	5.45	1.37	2.11	0.84	0.22
Food and Entertaining	SwinBERT		8.15	24.23	9.74	11.62	0.86	0.25
Home and Garden	SwinBERT		3.23	6.35	1.34	1.49	0.84	0.21
Cars and Other Vehicles	SwinBERT		3.12	4.51	0.17	2.07	0.84	0.20
Holidays and Traditions	SwinBERT		4.84	10.55	3.26	2.28	0.84	0.22
Education and Communications	SwinBERT		3.16	7.96	2.58	3.41	0.83	0.24
Personal Care and Style	SwinBERT		4.48	16.05	4.83	4.69	0.84	0.22
Pets and Animals	SwinBERT		2.62	7.17	0	6.25	0.83	0.21
Health	SwinBERT		2.68	6.75	0.33	0	0.83	0.19
Family Life	SwinBERT		2.46	9.78	0	14.29	0.83	0.20
Arts and Entertainment	SwinBERT		1.22	8.09	0	0	0.84	0.20
Sports and Fitness	SwinBERT		1.87	3.59	0	6.82	0.84	0.23
All	SwinBERT		5.12	13.31	4.65	5.86	0.85	0.23
Hobbies and Crafts	SwinBERT	✓	4.54	13.82	4.88	38.95	0.86	0.23
Food and Entertaining	SwinBERT	✓	8.08	37.64	9.82	32.60	0.87	0.24
Home and Garden	SwinBERT	✓	4.84	18.04	5.26	41.04	0.86	0.22
Cars and Other Vehicles	SwinBERT	✓	5.49	21.59	6.64	27.80	0.87	0.22
Holidays and Traditions	SwinBERT	✓	5.52	20.56	4.29	32.42	0.86	0.23
Education and Communications	SwinBERT	✓	3.45	10.51	1.33	23.86	0.85	0.24
Personal Care and Style	SwinBERT	✓	5.86	25.79	6.25	39.84	0.86	0.22
Pets and Animals	SwinBERT	✓	4.59	18.74	0	21.25	0.86	0.22
Health	SwinBERT	✓	2.27	5.06	0	10.91	0.85	0.20
Family Life	SwinBERT	✓	4.96	9.52	3.57	42.86	0.85	0.22
Arts and Entertainment	SwinBERT	✓	2.90	13.16	0	11.11	0.85	0.21
Sports and Fitness	SwinBERT	✓	2.65	12.78	0.91	54.55	0.86	0.24
Hobbies and Crafts	Joint (Ours)	✓	3.98	18.26	3.61	35.31	0.86	0.23
Food and Entertaining	Joint (Ours)	✓	4.22	30.35	2.63	57.67	0.86	0.23
Home and Garden	Joint (Ours)	✓	4.24	14.88	4.43	34.33	0.86	0.22
Cars and Other Vehicles	Joint (Ours)	✓	5.41	22.20	6.65	28.33	0.87	0.23
Holidays and Traditions	Joint (Ours)	✓	4.40	22.83	3.48	38.81	0.85	0.22
Education and Communications	Joint (Ours)	✓	4.01	19.37	4.17	27.27	0.85	0.23
Personal Care and Style	Joint (Ours)	✓	3.37	18.09	4.74	57.03	0.85	0.23
Pets and Animals	Joint (Ours)	✓	3.55	13.99	3.12	12.50	0.85	0.23
Health	Joint (Ours)	✓	2.33	11.37	2.42	30.91	0.85	0.19
Family Life	Joint (Ours)	✓	3.68	3.41	7.14	35.71	0.84	0.22
Arts and Entertainment	Joint (Ours)	✓	1.67	0	10.00	11.11	0.84	0.21
Sports and Fitness	Joint (Ours)	✓	2.54	17.79	2.65	40.91	0.86	0.23
All	SwinBERT	✓	5.94	24.66	6.67	35.09	0.86	0.23
All	Joint (Ours)	✓	4.13	23.01	3.54	43.88	0.86	0.23

Table 13. Step captioning results per video category on our HiREST test split. *FT: Finetuning on* HiREST.

Step Duration	Model	FT	METEOR	CIDEr	SPICE	Entailment (%)	BERTScore	CLIPScore
$<$ 8 secs	SwinBERT		5.73	16.72	5.40	6.74	0.84	0.23
8 - 18 secs	SwinBERT		5.18	13.95	4.95	6.05	0.85	0.23
$>$ 18 secs	SwinBERT		4.57	10.66	3.66	4.96	0.84	0.23
All	SwinBERT		5.12	13.31	4.65	5.86	0.85	0.23
$<$ 8 secs	SwinBERT	✓	6.25	25.32	6.94	25.37	0.86	0.23
8 - 18 secs	SwinBERT	✓	6.21	25.92	6.31	32.99	0.86	0.23
$>$ 18 secs	SwinBERT	✓	5.40	23.64	6.83	37.03	0.86	0.23
$<$ 8 secs	Joint (Ours)	✓	4.22	22.49	3.24	48.67	0.85	0.22
8 - 18 secs	Joint (Ours)	✓	4.02	22.55	3.36	41.25	0.86	0.23
$>$ 18 secs	Joint (Ours)	✓	4.17	24.83	4.02	41.29	0.86	0.22
All	SwinBERT	✓	5.94	24.66	6.67	35.09	0.86	0.23
All	Joint (Ours)	✓	4.13	23.01	3.54	43.88	0.86	0.23

Table 14. Step captioning results for various step durations on our HiREST test split. *FT: Finetuning on* HiREST.