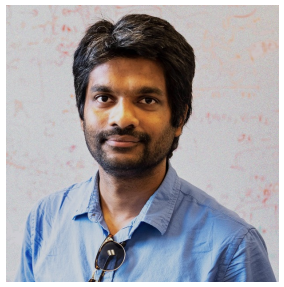
Hierarchical Video-Moment Retrieval and Step-Captioning

Abhay Zala[*1] Jaemin Cho[*1] Satwik Kottur[2] Xilun Chen[2] Barlas Oğuz[2] Yashar Mehdad[2] Mohit Bansal[1]









[1]UNC Chapel Hill, [2]Meta AI

hirest-cvpr2023.github.io

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

CVPR Paper Tag
THU-PM-234

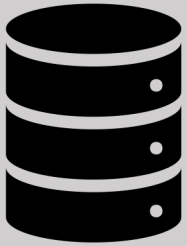

HOW TO
BUDGET FOR
A CAR
8:09


MANHATTAN
MOJITO
AMERICANO
WHISKEY SO
37:00

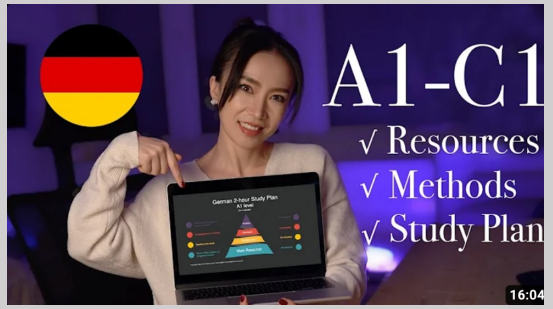

ALONE ACROSS AN OCEAN
& THE LOSS OF A RUDDER 1000 MILES FROM HAWAII
1:04:28


MUSIC PRODUCTION
IN
12 DAYS!
9:49

Videos



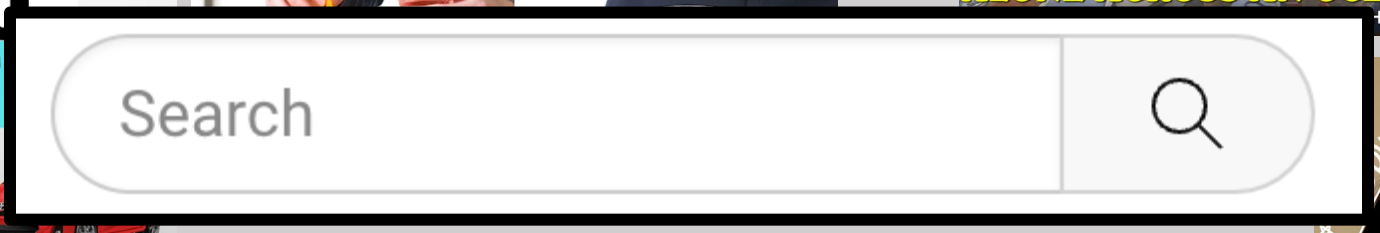

How to
Handstand
+
Hold it
longer
3:47


A1-C1
√ Resources
√ Methods
√ Study Plan
16:04


1500+
SAT SCORE
IN ONE WEEK
THE OFFICIAL
SAT
Study Guide
7:39


THE LEGEND OF
ZELDA
BREATH OF THE WILD
9 TIPS FOR
BEGINNERS
4:53


15-MINUTE
KOREAN RICE BOWLS
40:12



How to search useful information from large video corpus? 🤔

1. **Find a relevant video**

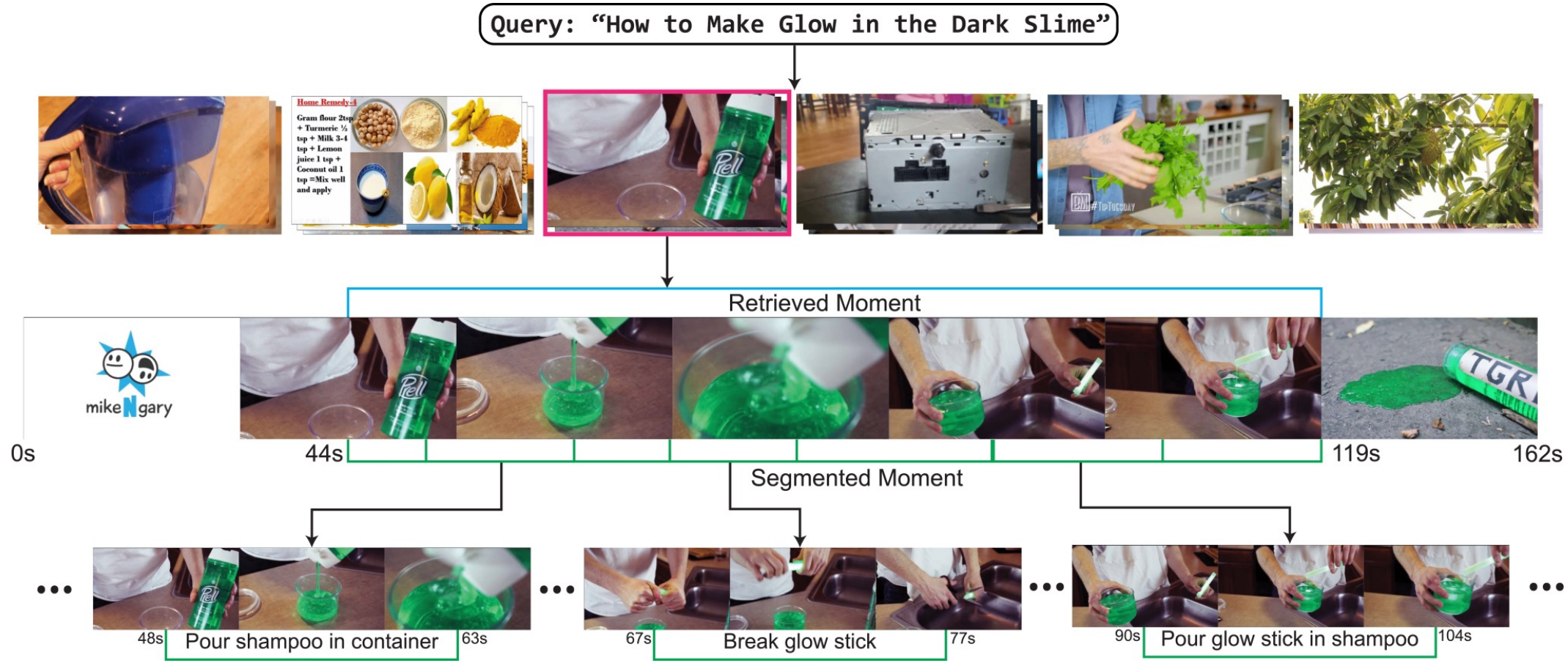
2. **Localize a relevant moment**

3. **Summarize**

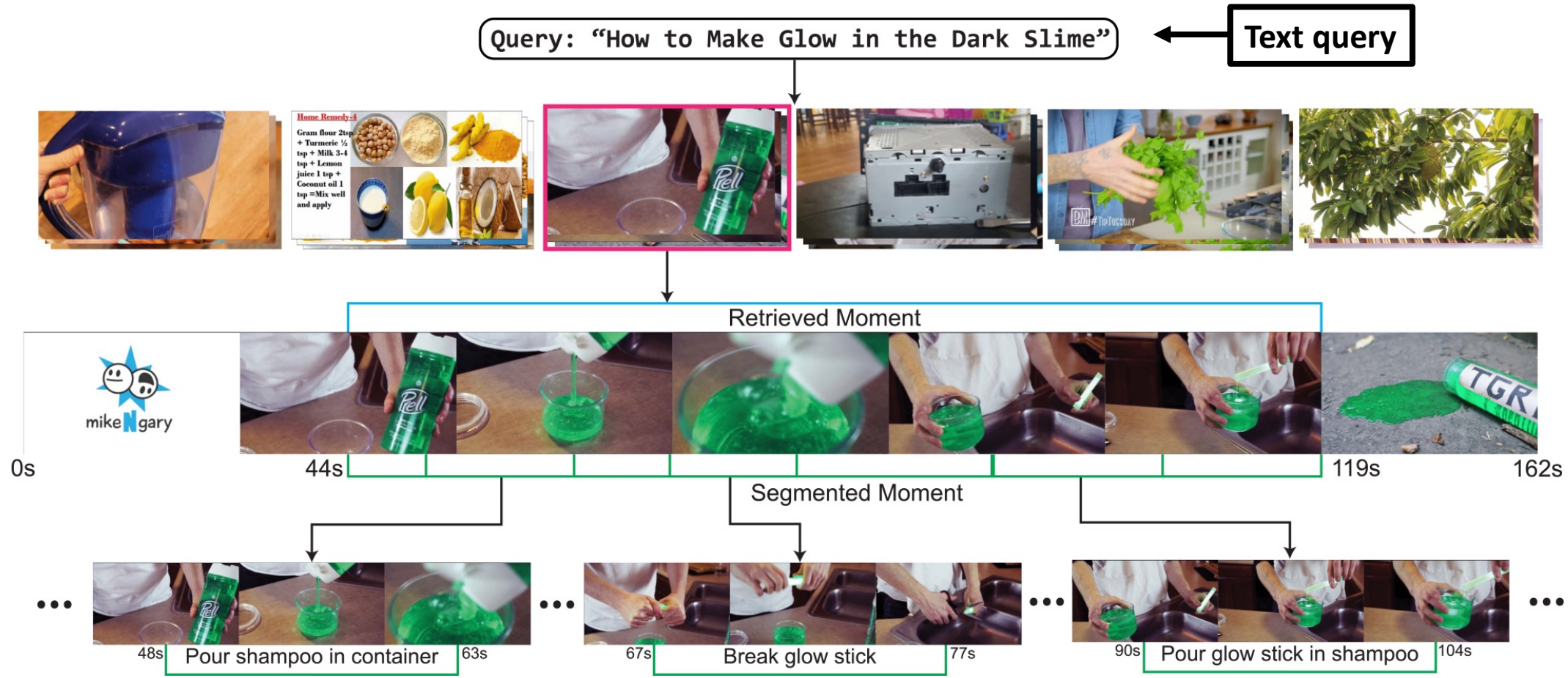
Previous work focuses on solving single task

This work studies all tasks in a hierarchical & end-to-end setup

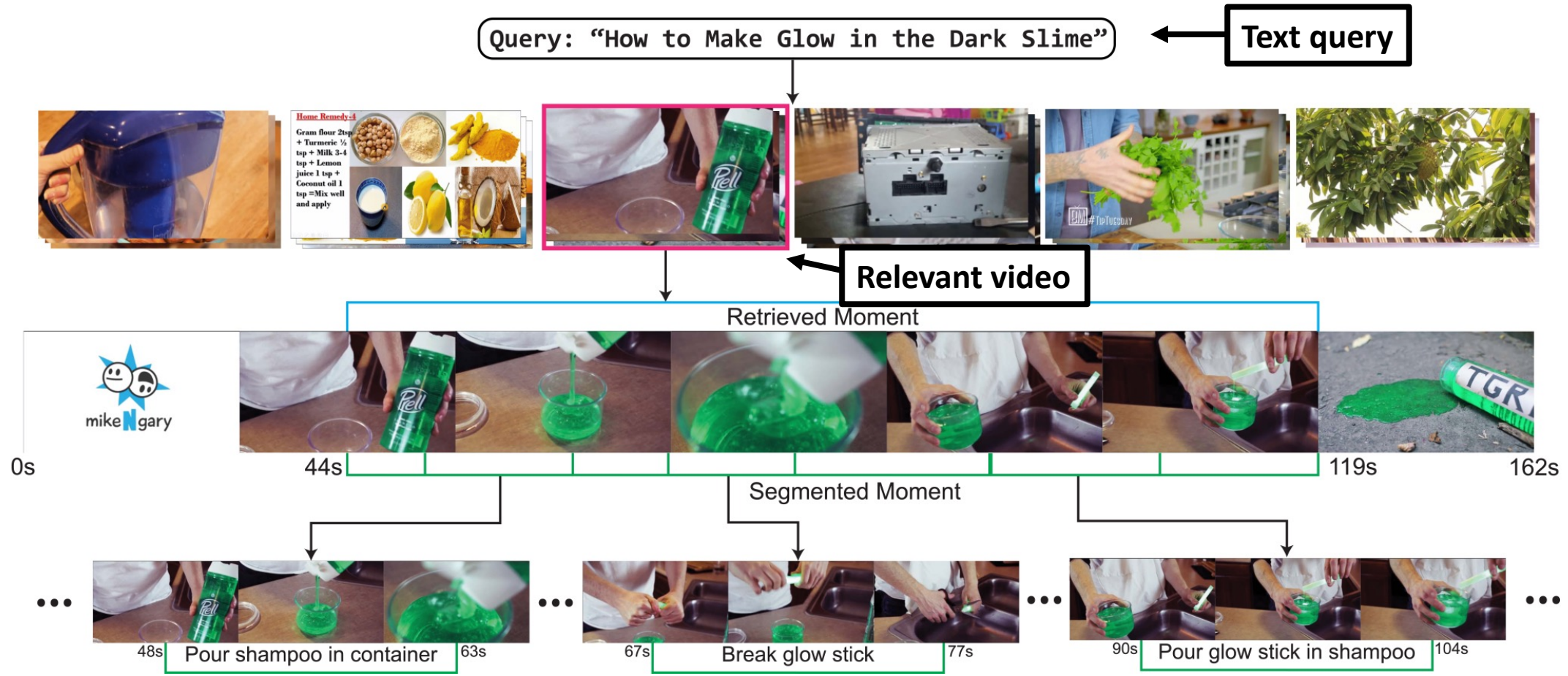
HiREST (HIerarchical RETrieval and STep-captioning)



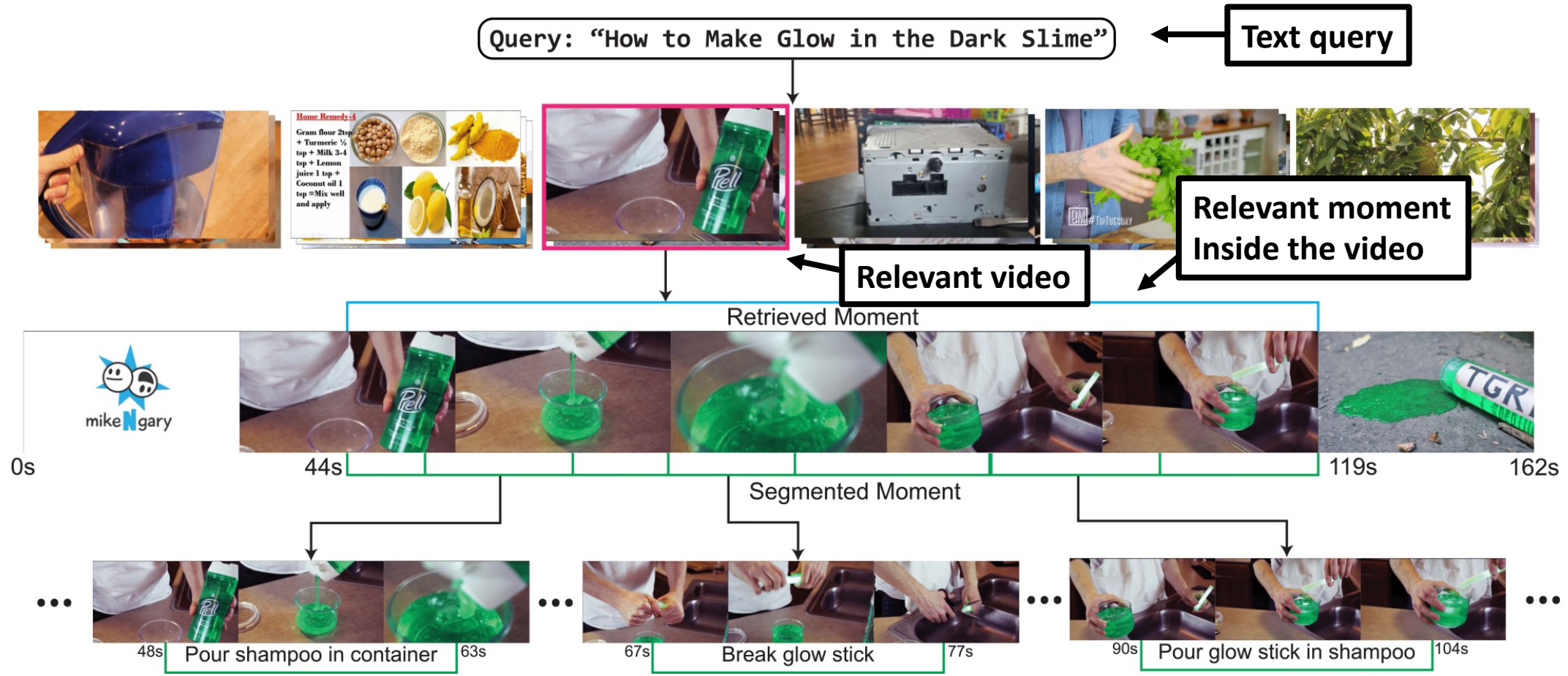
HiREST (HIerarchical RETrieval and STep-captioning)



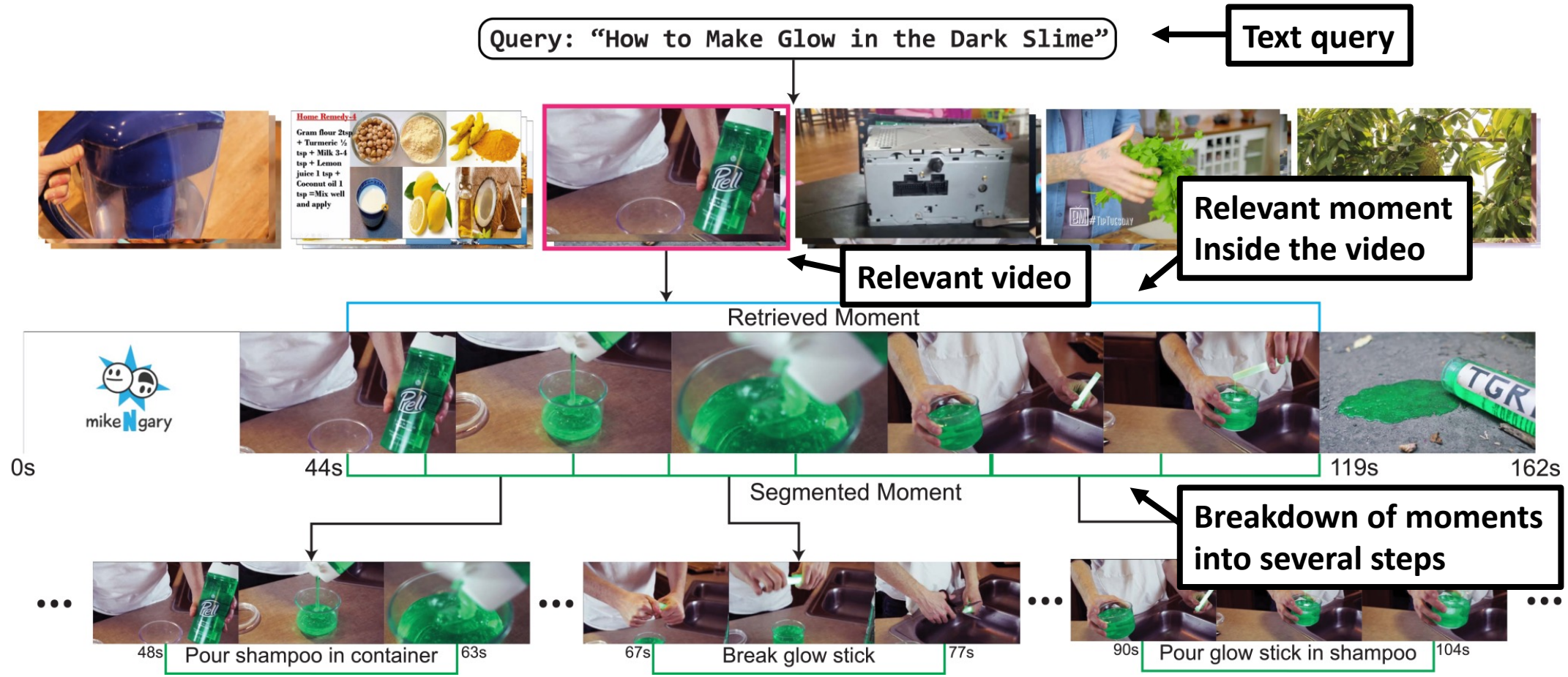
HiREST (HIerarchical RETrieval and STep-captioning)



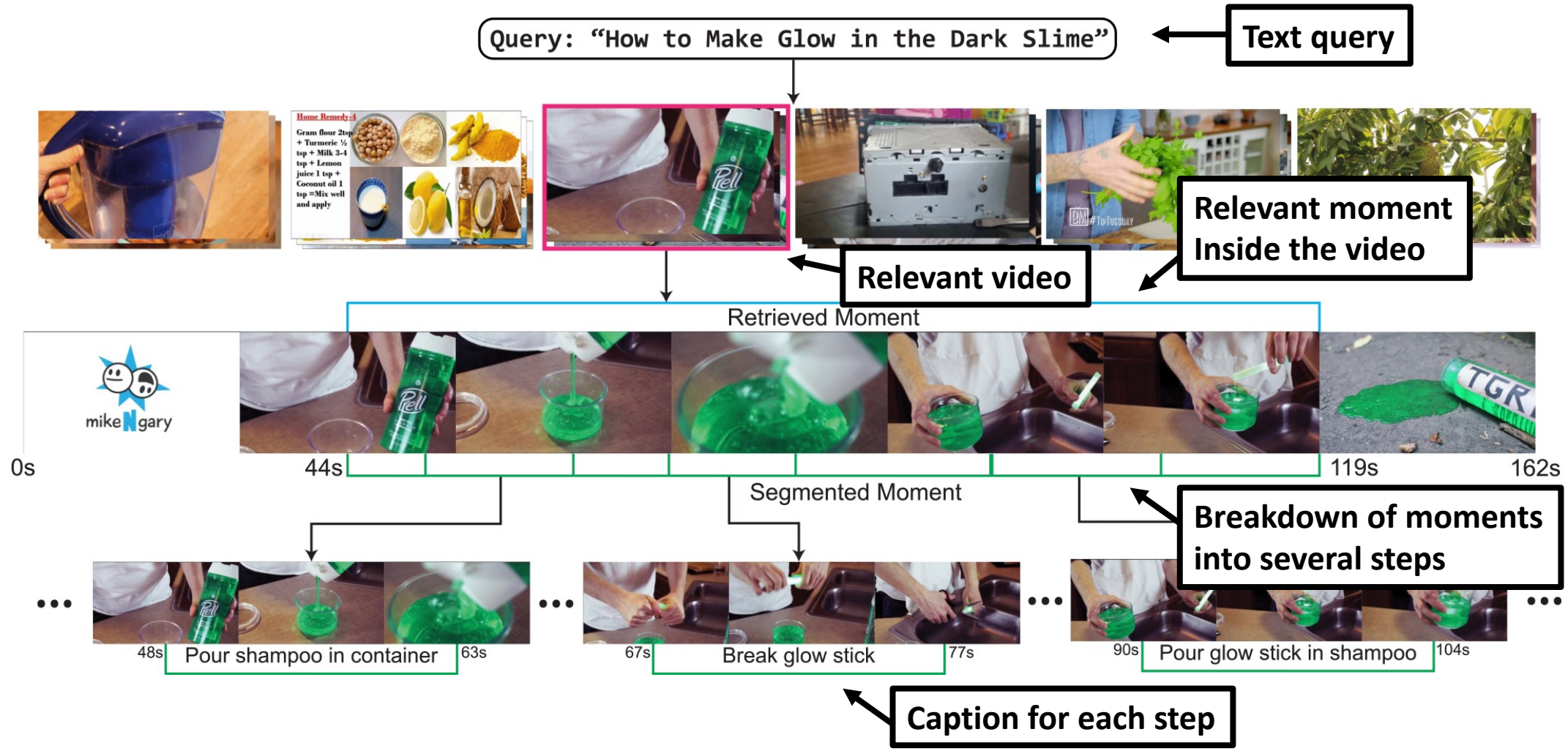
HiREST (HIerarchical RETrieval and STep-captioning)



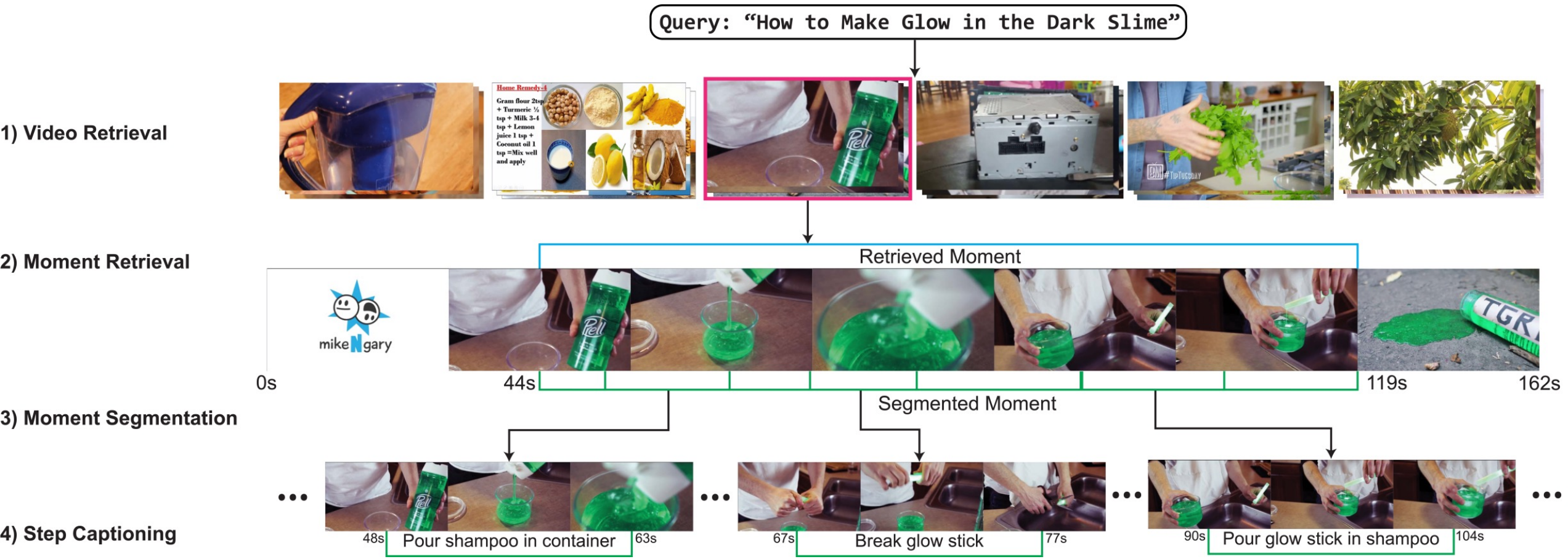
HiREST (HIerarchical RETrieval and STep-captioning)



HiREST (HIerarchical RETrieval and STep-captioning)



HiREST (HIerarchical RETrieval and STep-captioning)



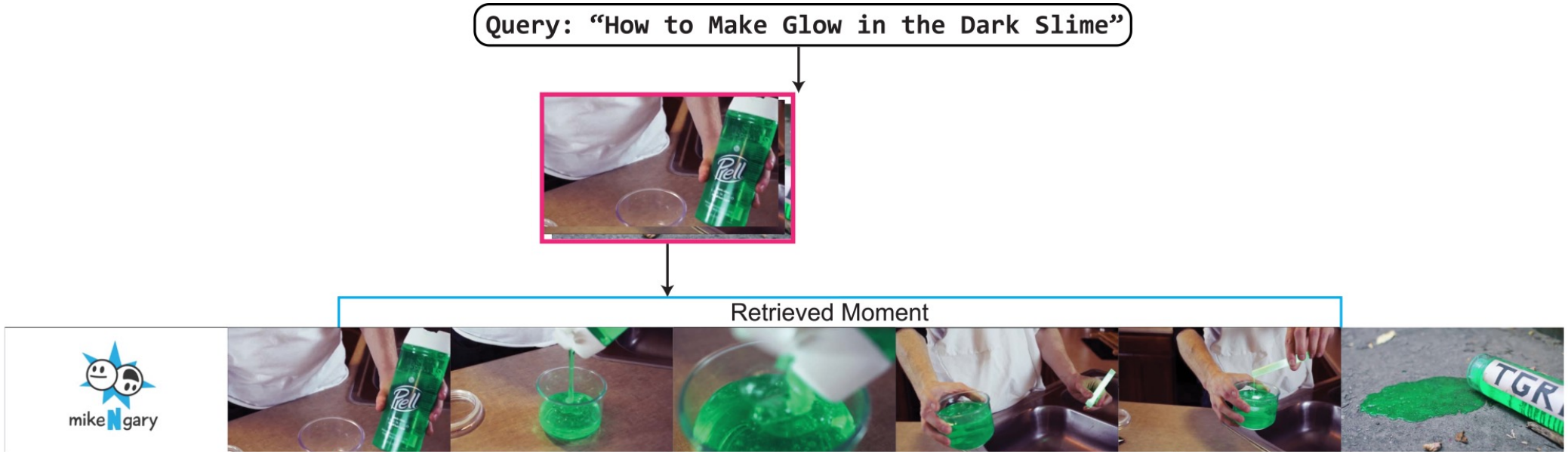
HiREST (HIerarchical RETrieval and STep-captioning)

Query: “How to Make Glow in the Dark Slime”

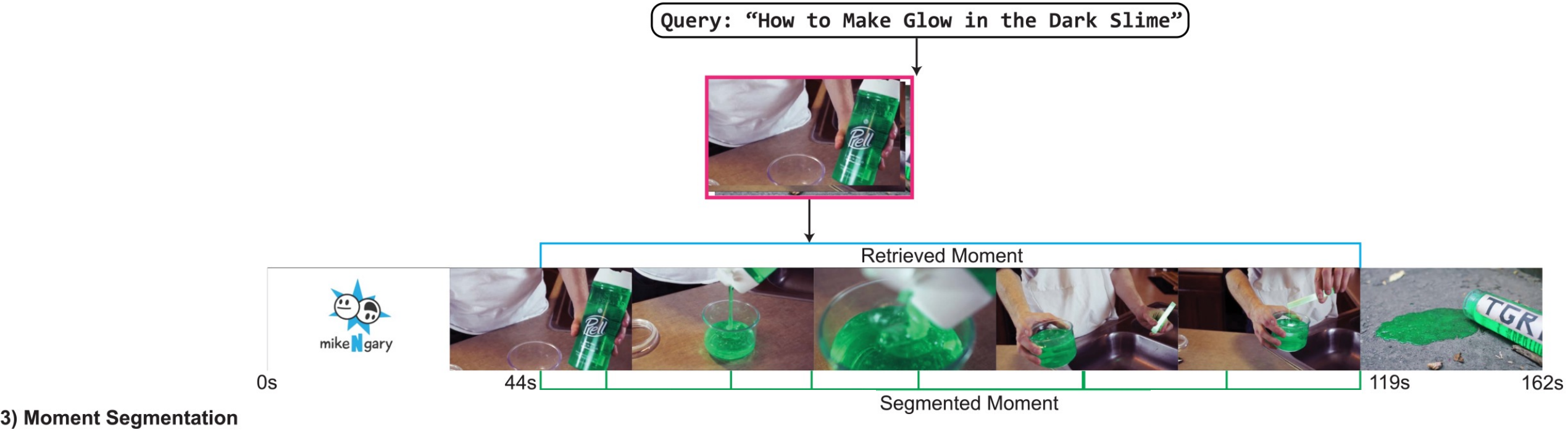
1) Video Retrieval

HiREST (**HI**erarchical **RE**trieval and **ST**ep-captioning)

2) Moment Retrieval

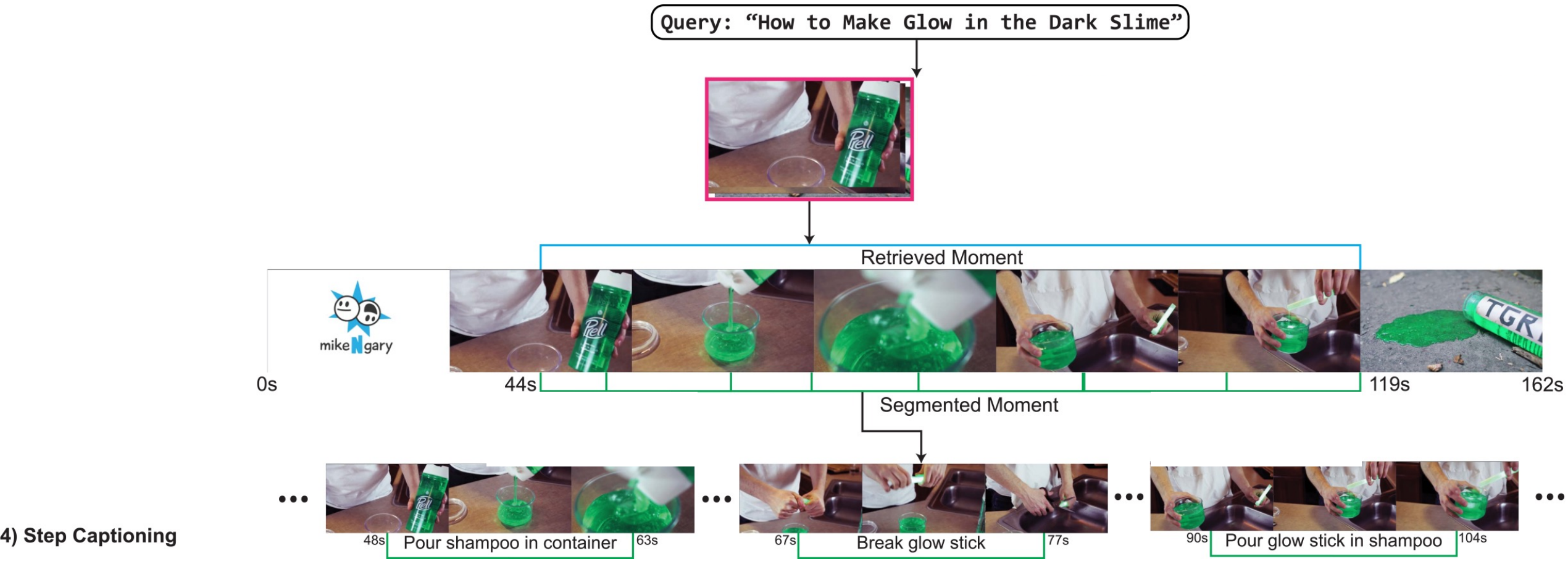


HiREST (**HI**erarchical **RE**trieval and **ST**ep-captioning)

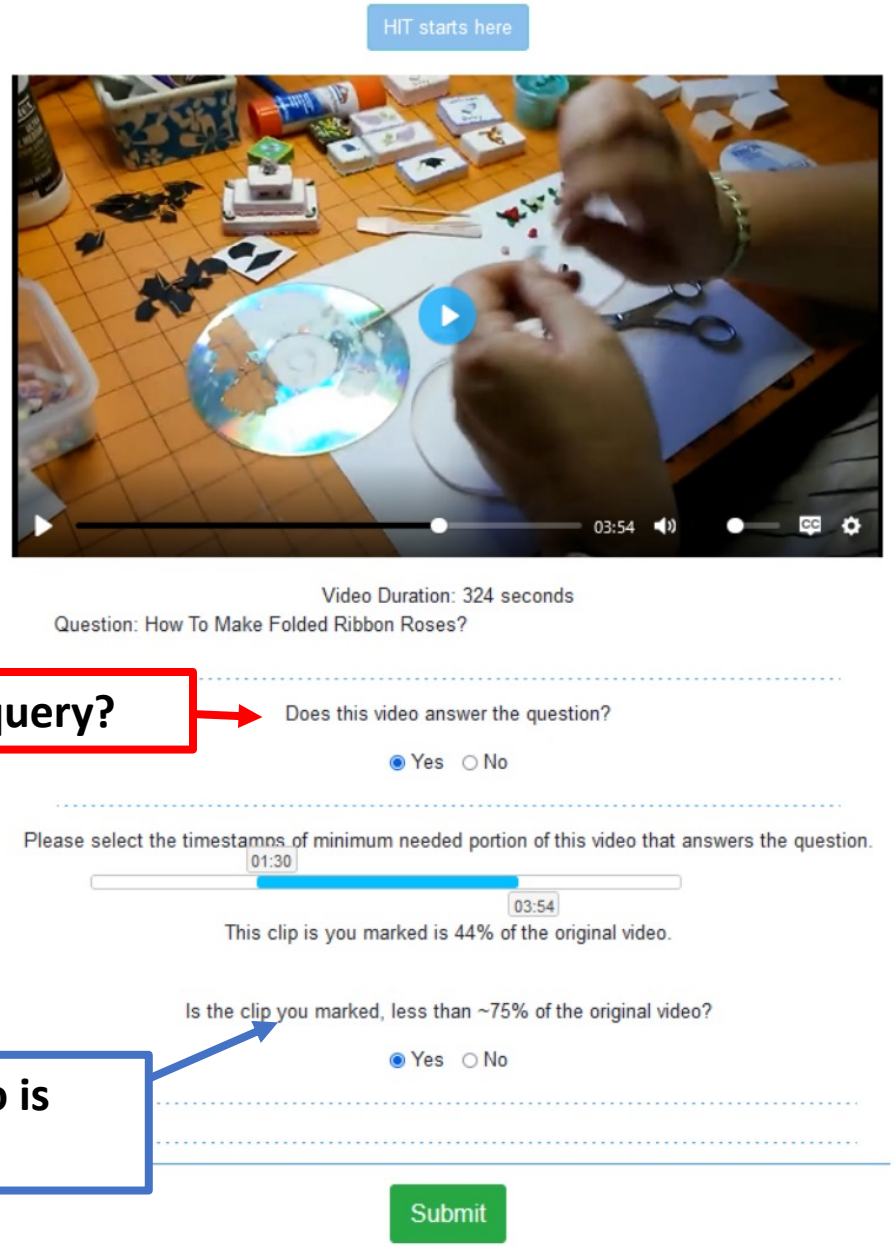


3) Moment Segmentation

HiREST (HIerarchical RETrieval and STep-captioning)



Dataset Collection



Stage 1. Video Retrieval and Moment Retrieval

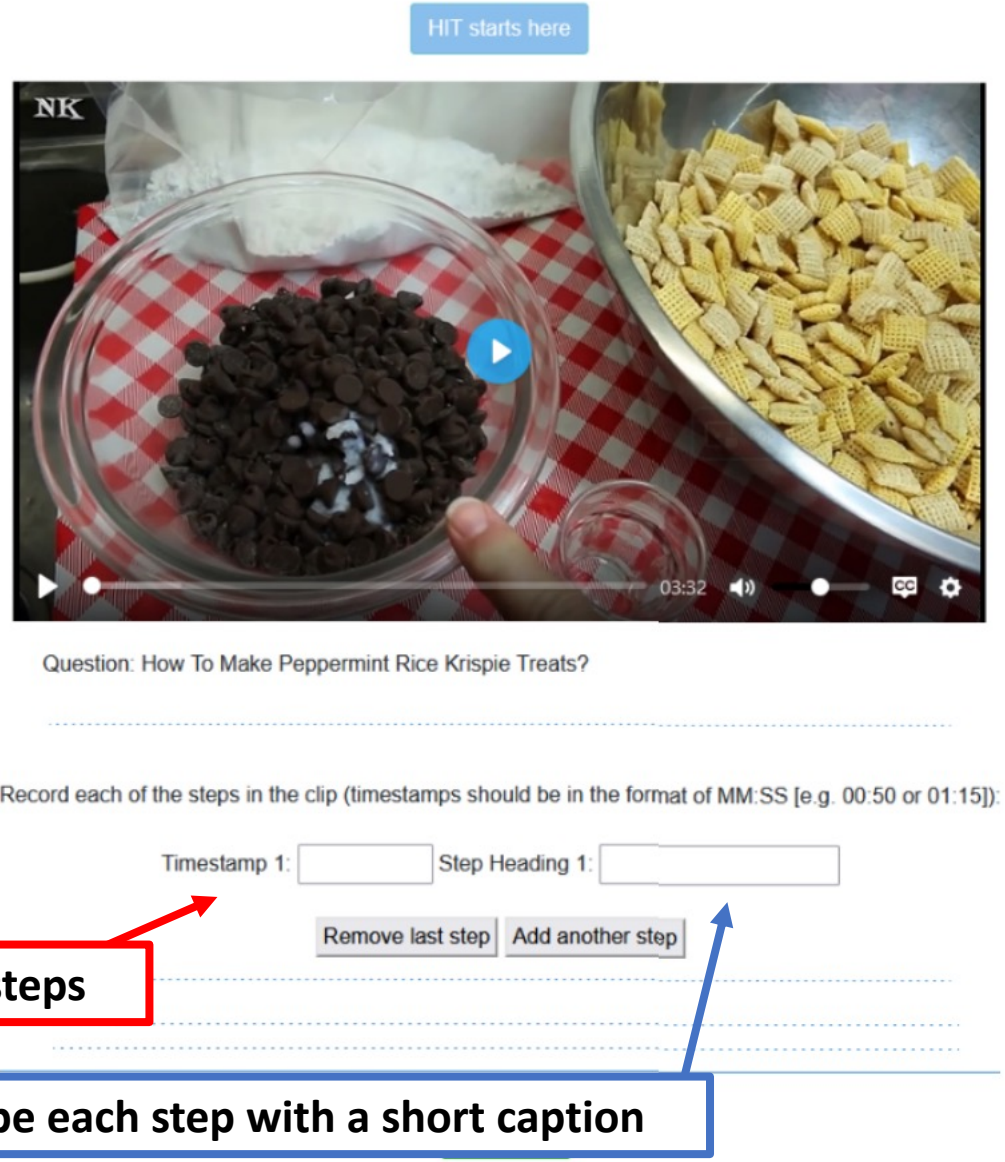
Is video relevant to text query?

Stage 2. Moment Segmentation and Step Captions

Which moment in the video is relevant to text query?

Dataset Collection

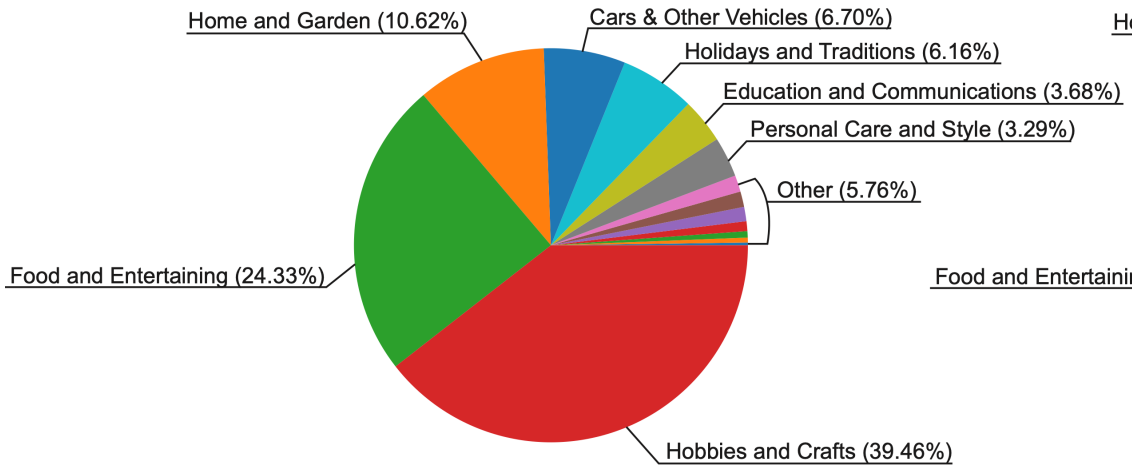
Stage 1. Video Retrieval and Moment Retrieval



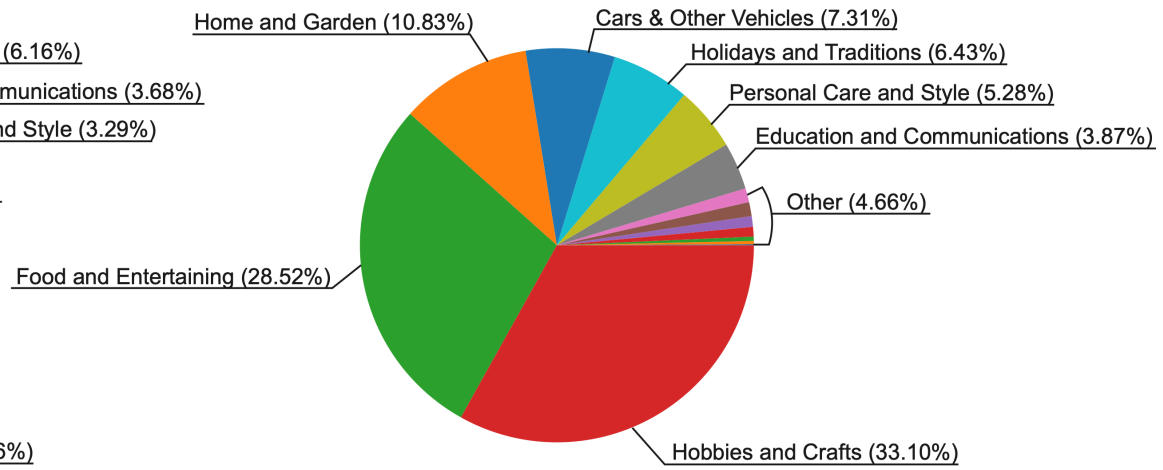
Stage 2. Moment Segmentation and Step Captions

Categories of Text Queries

wide variety of categories



(a) all videos



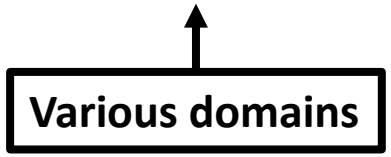
(b) videos with step caption annotations

HiREST vs. Other Video Datasets

Dataset	Domain	Step caption	# Videos / # Steps	# Steps per Moment	# Words per Caption	# Unique Captions	Avg. Duration (s) Video / Step
COIN [36]	Open	Predefined steps	11.8K / 46K	3.9	4.8	0.8K	142 / 14.9
CrossTask [46]	Open	Predefined steps	4.7K / 21K	7.4	2.4	0.1K	297 / 9.6
YouCook2 [45]	Cooking	Manually written	2K / 14K	7.7	8.8	13K	316 / 19.7
HiREST (Ours)	Open	Manually written	3.4K (1.1K w/ steps) / 8.6K	7.6	4.4	7.9K	263 / 18.9

HiREST vs. Other Video Datasets

Dataset	Domain	Step caption	# Videos / # Steps	# Steps per Moment	# Words per Caption	# Unique Captions	Avg. Duration (s) Video / Step
COIN [36]	Open	Predefined steps	11.8K / 46K	3.9	4.8	0.8K	142 / 14.9
CrossTask [46]	Open	Predefined steps	4.7K / 21K	7.4	2.4	0.1K	297 / 9.6
YouCook2 [45]	Cooking	Manually written	2K / 14K	7.7	8.8	13K	316 / 19.7
HiREST (Ours)	Open	Manually written	3.4K (1.1K w/ steps) / 8.6K	7.6	4.4	7.9K	263 / 18.9



HiREST vs. Other Video Datasets

Dataset	Domain	Step caption	# Videos / # Steps	# Steps per Moment	# Words per Caption	# Unique Captions	Avg. Duration (s) Video / Step
COIN [36]	Open	Predefined steps	11.8K / 46K	3.9	4.8	0.8K	142 / 14.9
CrossTask [46]	Open	Predefined steps	4.7K / 21K	7.4	2.4	0.1K	297 / 9.6
YouCook2 [45]	Cooking	Manually written	2K / 14K	7.7	8.8	13K	316 / 19.7
HiREST (Ours)	Open	Manually written	3.4K (1.1K w/ steps) / 8.6K	7.6	4.4	7.9K	263 / 18.9

Various domains

High-quality, diverse step captions written by human annotators

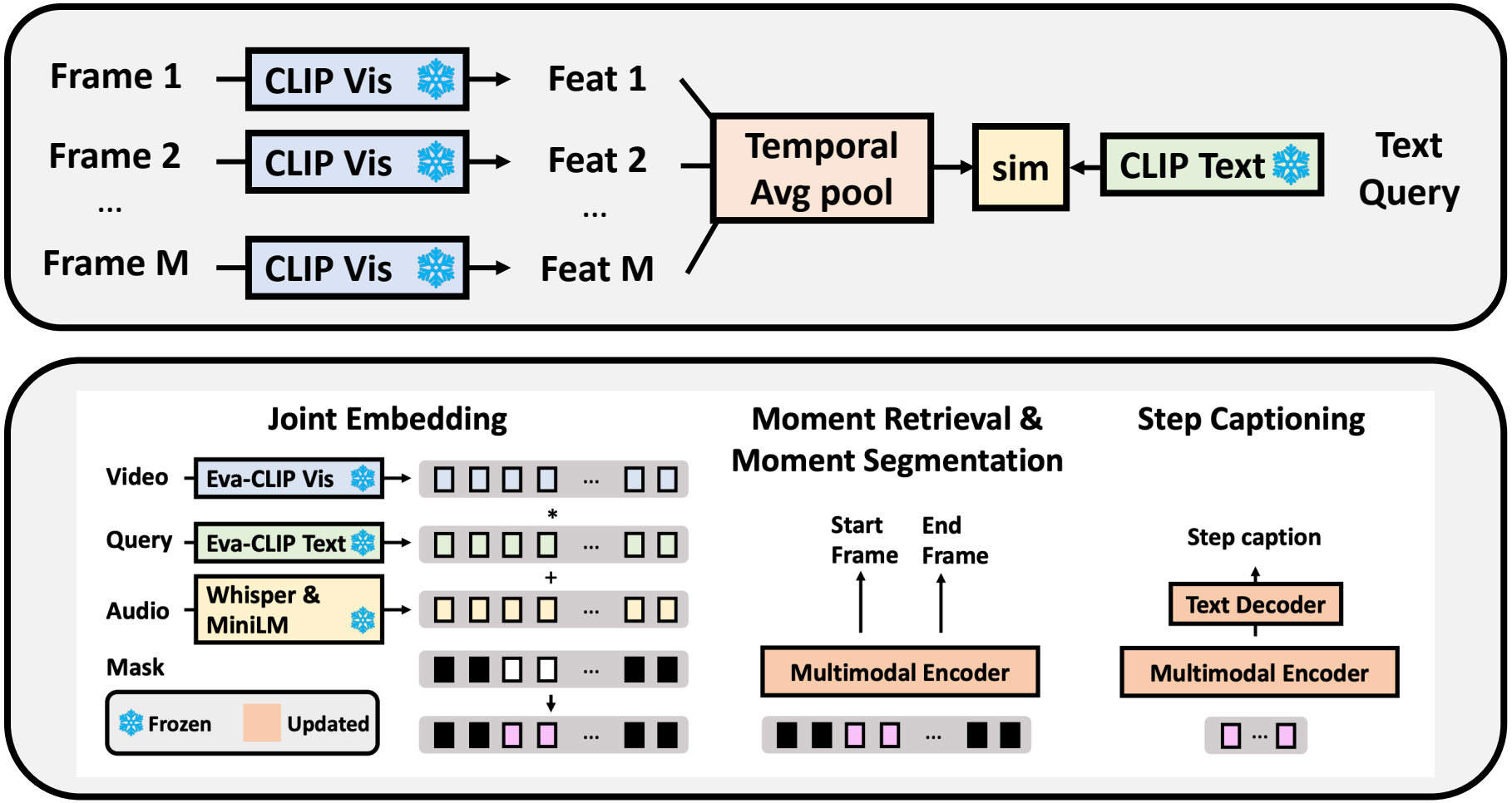
Baseline Model for HiREST

Video Retrieval

Moment Retrieval

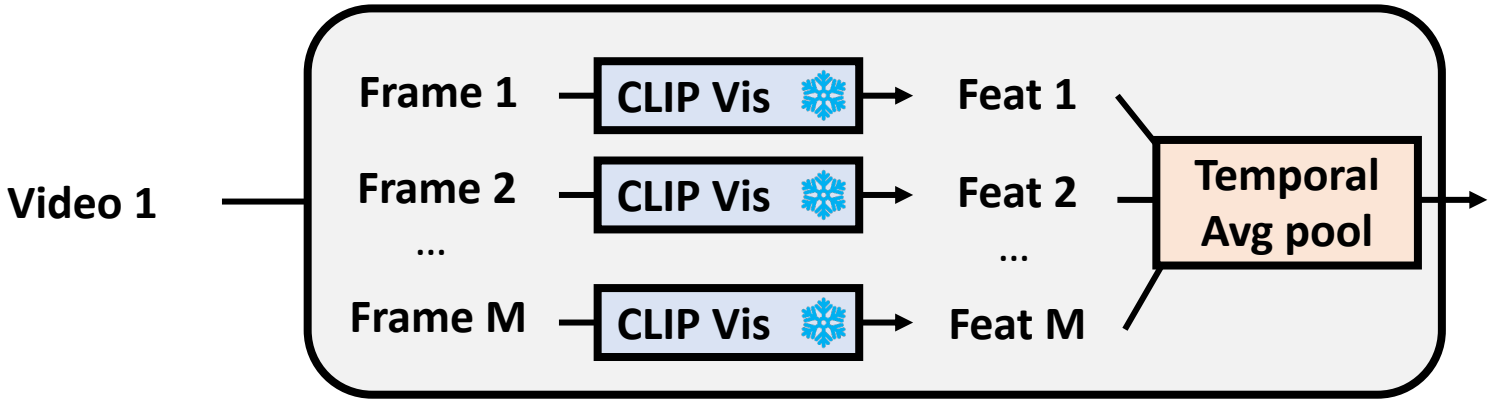
Moment Segmentation

Step Captioning



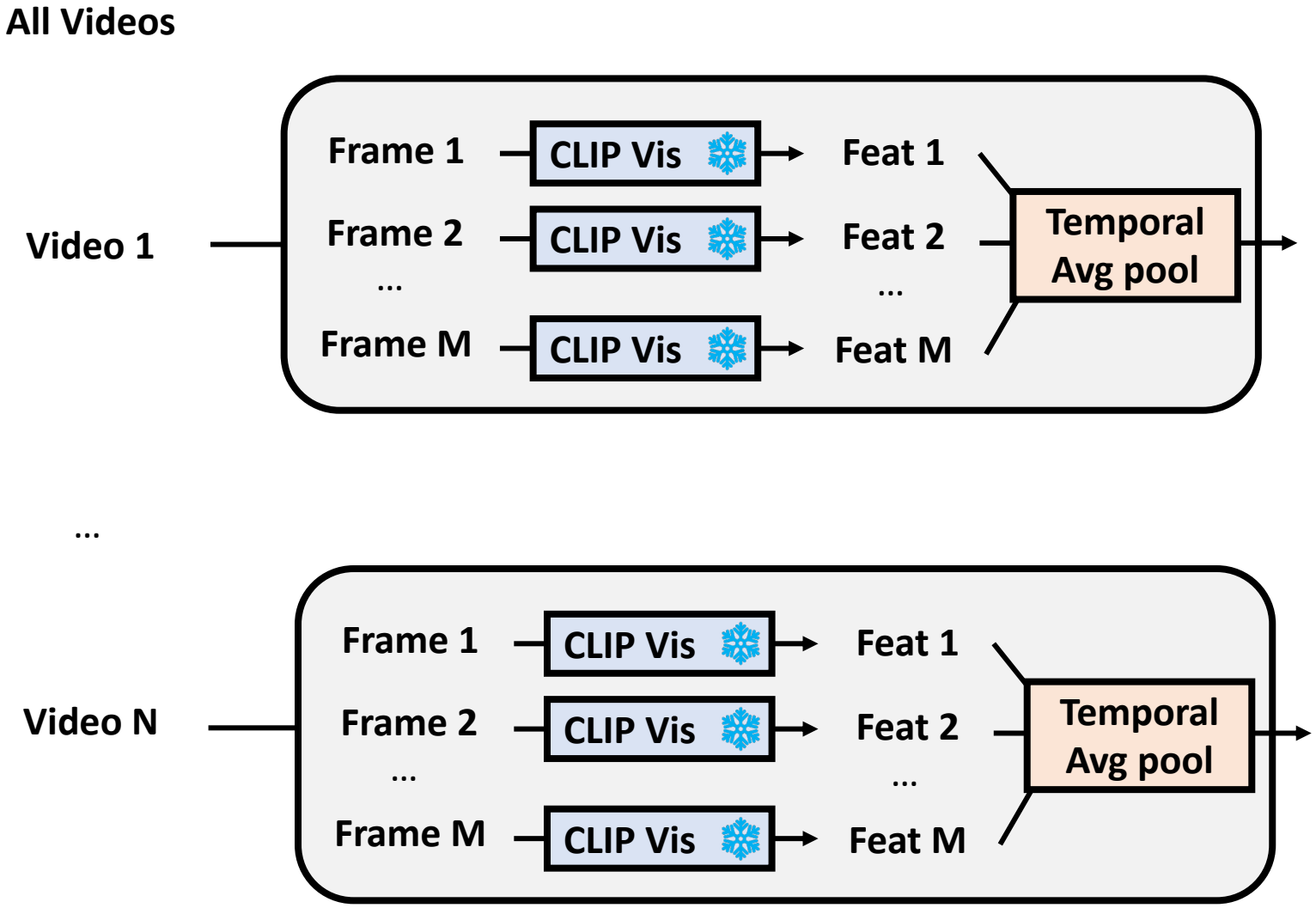
Baseline Model for HiREST – Video Retrieval

Find the right video



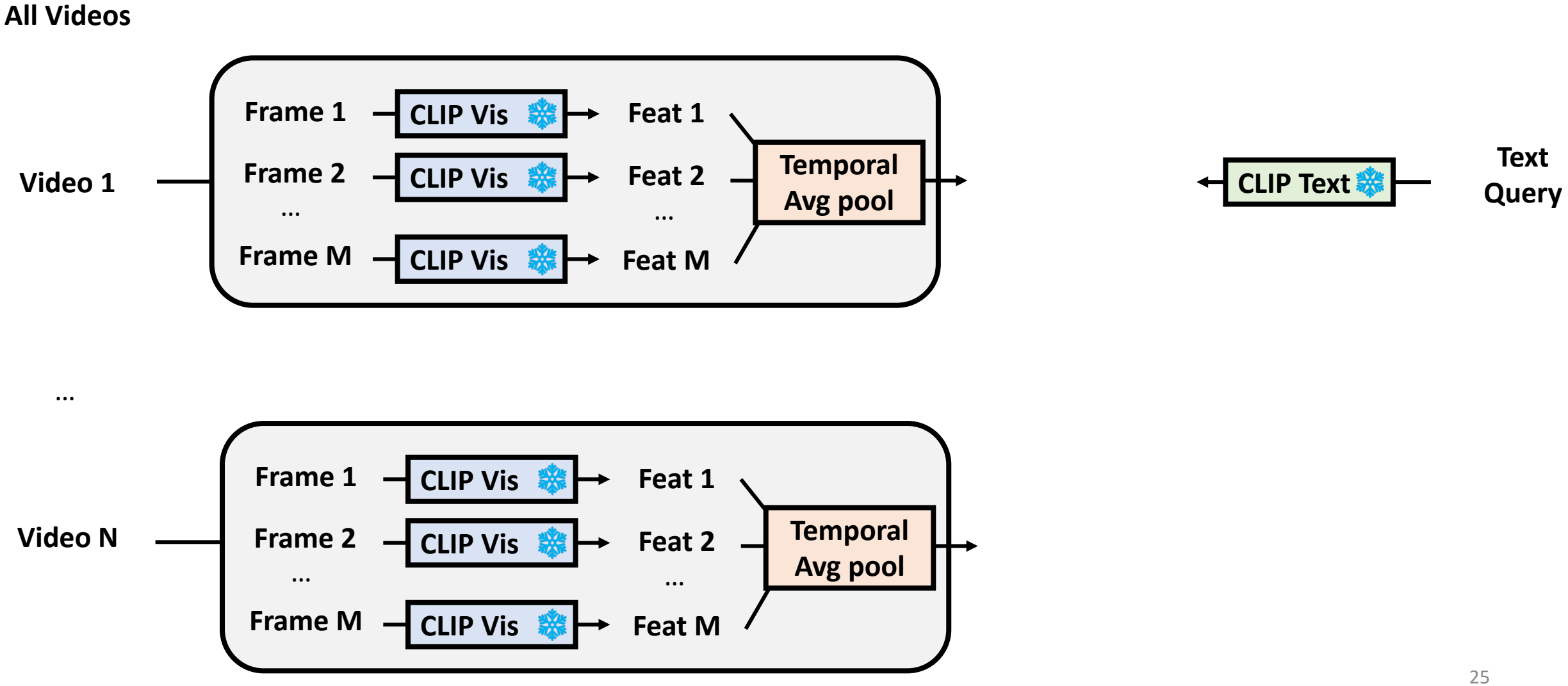
Baseline Model for HiREST – Video Retrieval

Find the right video



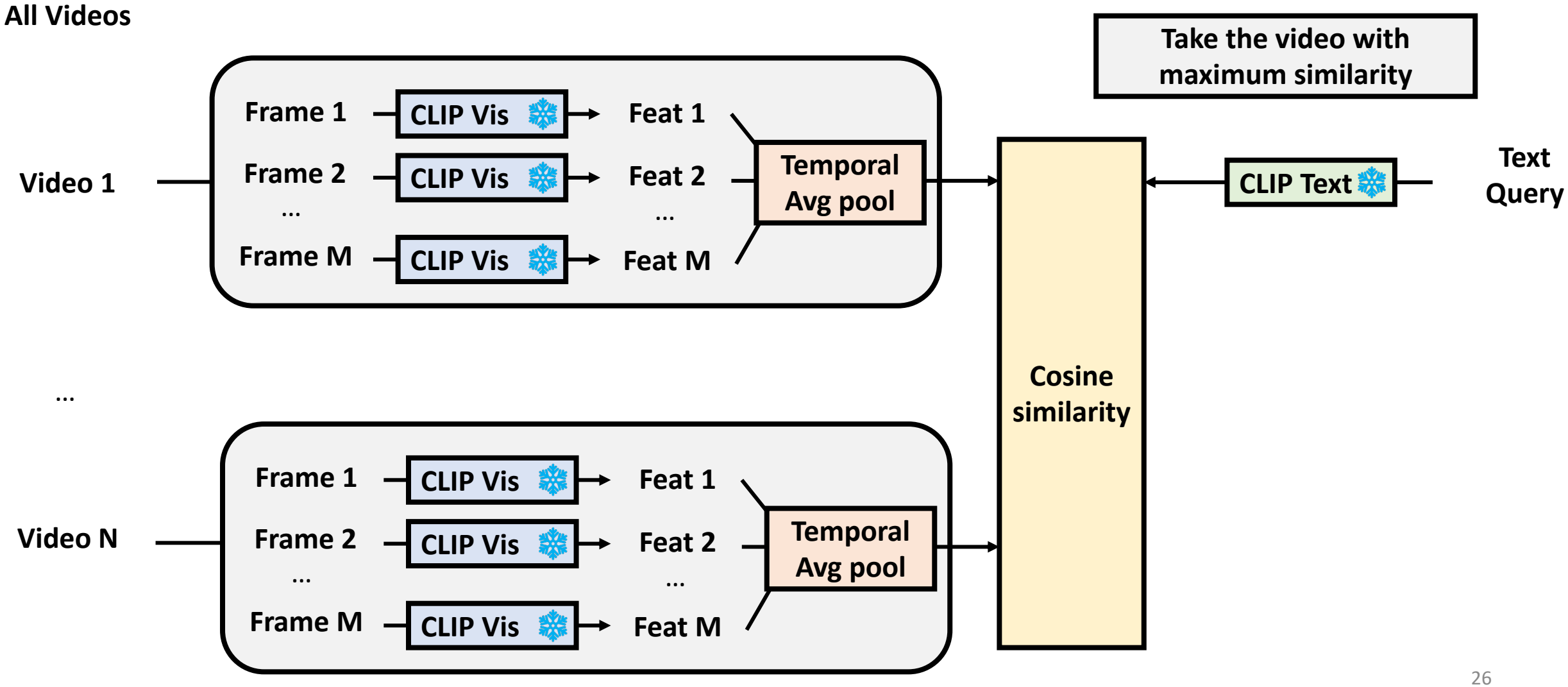
Baseline Model for HiREST – Video Retrieval

Find the right video



Baseline Model for HiREST – Video Retrieval

Find the right video



Baseline Model for HiREST – Video Retrieval

Find the right video

Image Models (CLIP-B) vs. Video Models (Frozen-in-Time / MIL-NCE)

- Although CLIP has not seen videos, CLIP is as effective as video models

Model	Frames	FT	R@1	R@5	R@10
CLIP-B/32	1		11.4	20.7	27.3
CLIP-B/32	4		12.5	28.8	37.4
CLIP-B/32	10		13.0	31.7	39.9
CLIP-B/32	20		13.0	33.3	41.2
CLIP-B/32	32		12.6	33.0	41.8
Frozen-in-Time	4		7.0	19.4	26.7
MIL-NCE (S3D)	32		13.9	31.1	41.4
CLIP-B/32	1	✓	11.5	22.7	27.1
CLIP-B/32	4	✓	13.9	29.5	39.4
CLIP-B/32	10	✓	11.4	31.3	41.4
CLIP-B/32	20	✓	12.3	31.7	41.6
CLIP-B/32	32	✓	13.0	32.1	41.9
EVA-CLIP-G/14	1		18.9	32.6	37.5
EVA-CLIP-G/14	4		20.7	43.6	53.7
EVA-CLIP-G/14	10		26.0	48.5	58.8
EVA-CLIP-G/14	20		**26.4**	**51.1**	**61.5**
EVA-CLIP-G/14	32		26.0	50.0	61.4

Table 2. Video retrieval results on HIREST test split. CLIP/EVA-CLIP results are based on temporal average pooling. *FT: finetuning on* HIREST, *R@k: Recall@k.* MIL-NCE was trained on the HowTo100M dataset, which is the video source of HIREST.

Baseline Model for HiREST – Video Retrieval

Find the right video

CLIP-B zeroshot vs. CLIP-B Finetuning

- Finetuning CLIP does not help much

Model	Frames	FT	R@1	R@5	R@10
CLIP-B/32	1		11.4	20.7	27.3
CLIP-B/32	4		12.5	28.8	37.4
CLIP-B/32	10		13.0	31.7	39.9
CLIP-B/32	20		13.0	33.3	41.2
CLIP-B/32	32		12.6	33.0	41.8
Frozen-in-Time	4		7.0	19.4	26.7
MIL-NCE (S3D)	32		13.9	31.1	41.4
CLIP-B/32	1	✓	11.5	22.7	27.1
CLIP-B/32	4	✓	13.9	29.5	39.4
CLIP-B/32	10	✓	11.4	31.3	41.4
CLIP-B/32	20	✓	12.3	31.7	41.6
CLIP-B/32	32	✓	13.0	32.1	41.9
EVA-CLIP-G/14	1		18.9	32.6	37.5
EVA-CLIP-G/14	4		20.7	43.6	53.7
EVA-CLIP-G/14	10		26.0	48.5	58.8
EVA-CLIP-G/14	20		**26.4**	**51.1**	**61.5**
EVA-CLIP-G/14	32		26.0	50.0	61.4

Table 2. Video retrieval results on HiREST test split. CLIP/EVA-CLIP results are based on temporal average pooling. *FT: finetuning on* HiREST, *R@k: Recall@k.* MIL-NCE was trained on the HowTo100M dataset, which is the video source of HiREST.

Baseline Model for HiREST – Video Retrieval

Find the right video

EVA-CLIP-G vs. CLIP-B

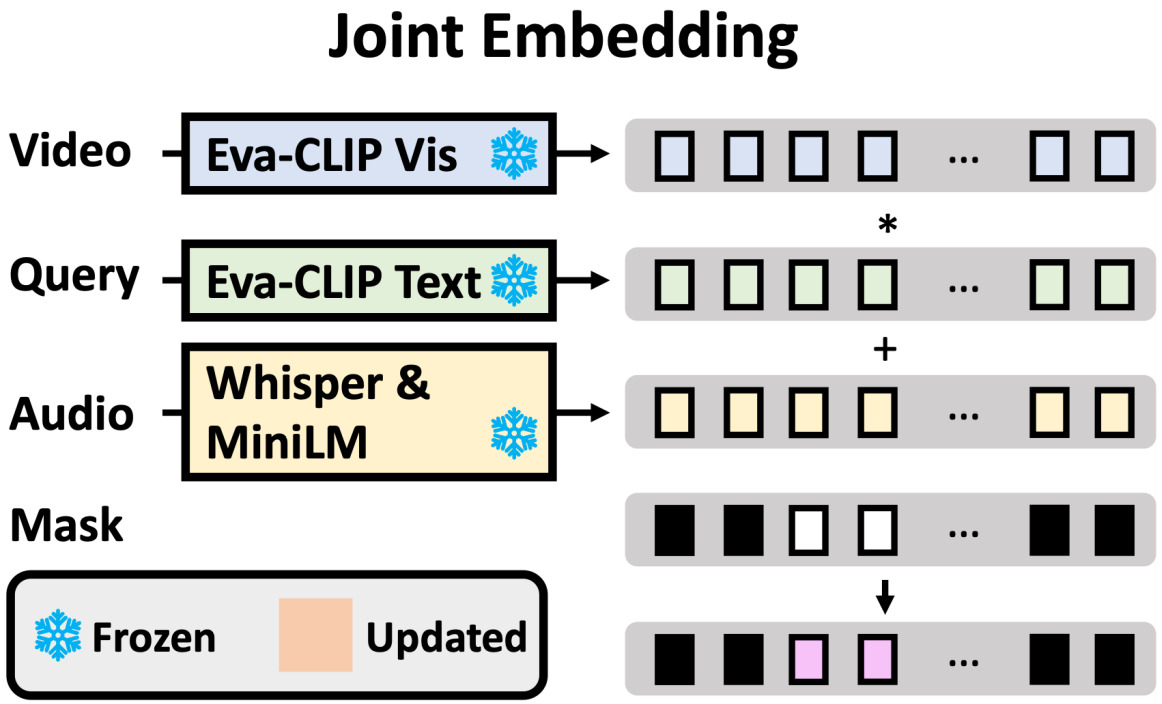
- Bigger model helps

Model	Frames	FT	R@1	R@5	R@10
CLIP-B/32	1		11.4	20.7	27.3
CLIP-B/32	4		12.5	28.8	37.4
CLIP-B/32	10		13.0	31.7	39.9
CLIP-B/32	20		13.0	33.3	41.2
CLIP-B/32	32		12.6	33.0	41.8
Frozen-in-Time	4		7.0	19.4	26.7
MIL-NCE (S3D)	32		13.9	31.1	41.4
CLIP-B/32	1	✓	11.5	22.7	27.1
CLIP-B/32	4	✓	13.9	29.5	39.4
CLIP-B/32	10	✓	11.4	31.3	41.4
CLIP-B/32	20	✓	12.3	31.7	41.6
CLIP-B/32	32	✓	13.0	32.1	41.9
EVA-CLIP-G/14	1		18.9	32.6	37.5
EVA-CLIP-G/14	4		20.7	43.6	53.7
EVA-CLIP-G/14	10		26.0	48.5	58.8
EVA-CLIP-G/14	20		**26.4**	**51.1**	**61.5**
EVA-CLIP-G/14	32		26.0	50.0	61.4

Table 2. Video retrieval results on HiREST test split. CLIP/EVA-CLIP results are based on temporal average pooling. *FT: finetuning on* HiREST, *R@k: Recall@k.* MIL-NCE was trained on the HowTo100M dataset, which is the video source of HiREST.

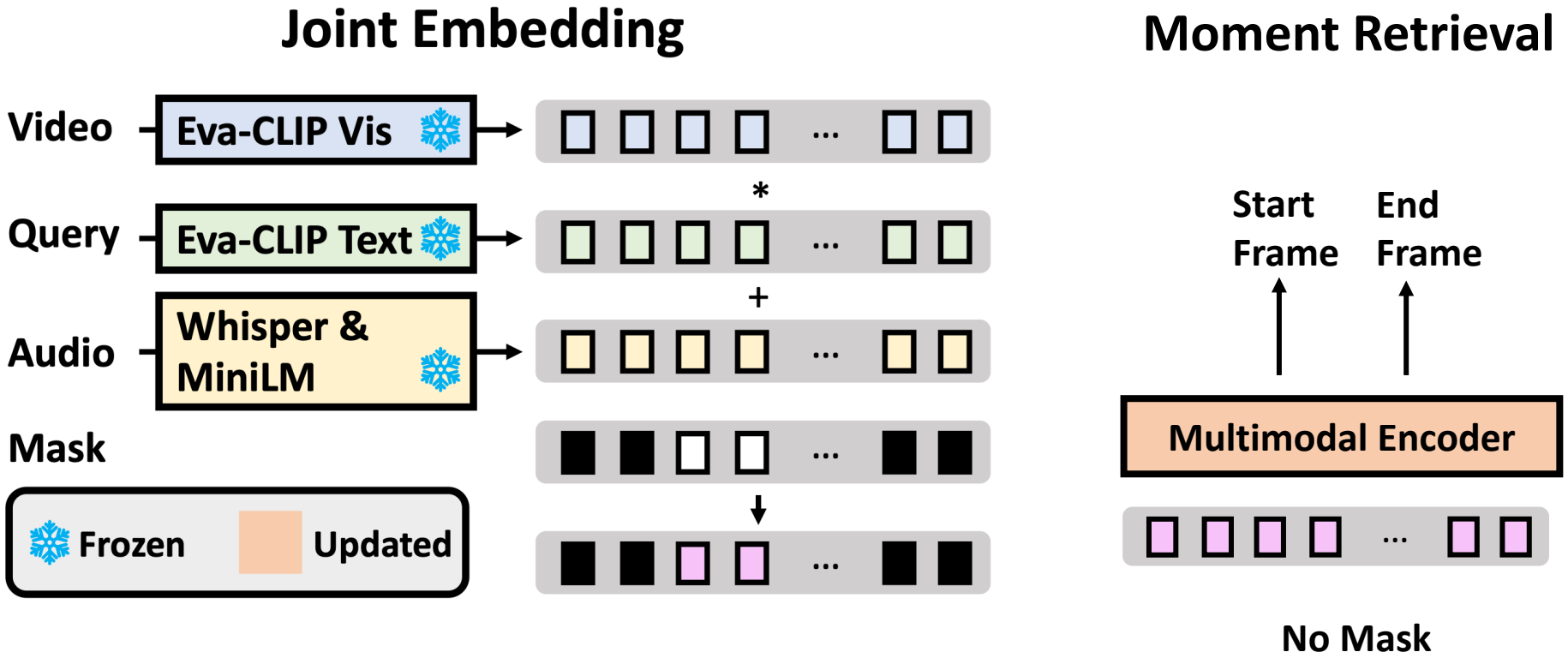
Baseline Model for HiREST – Joint Embedding

Combining multimodal information of video, text query, and audio (ASR)



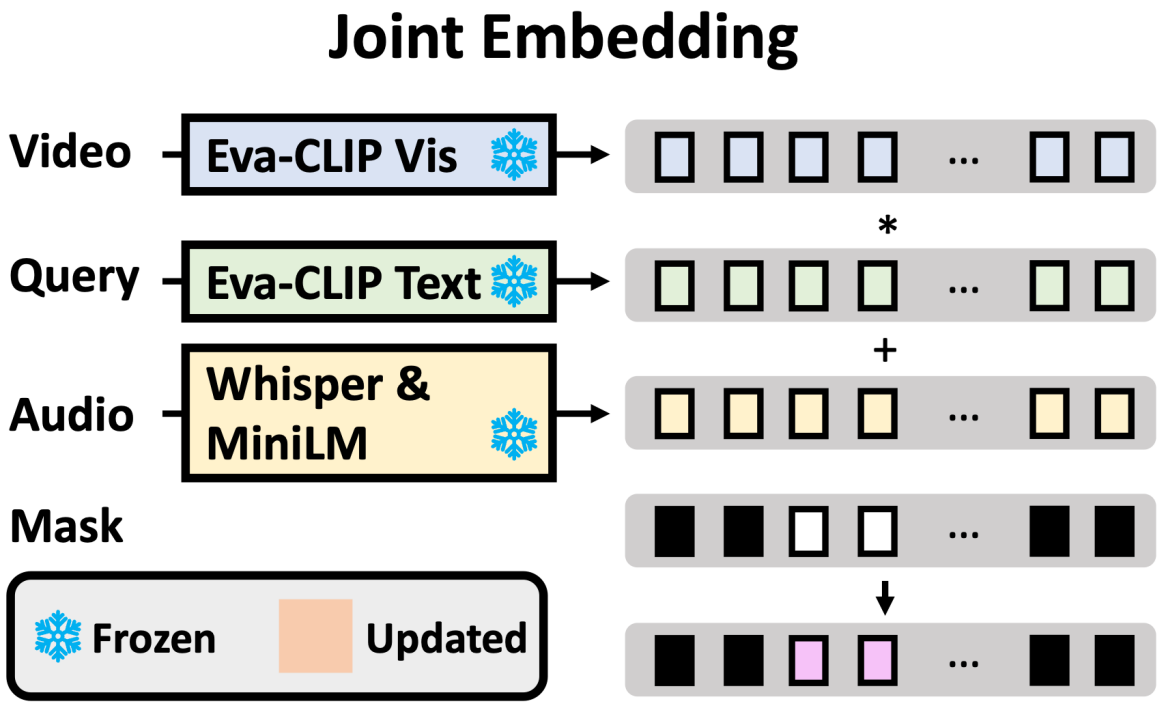
Baseline Model for HiREST – Moment Retrieval

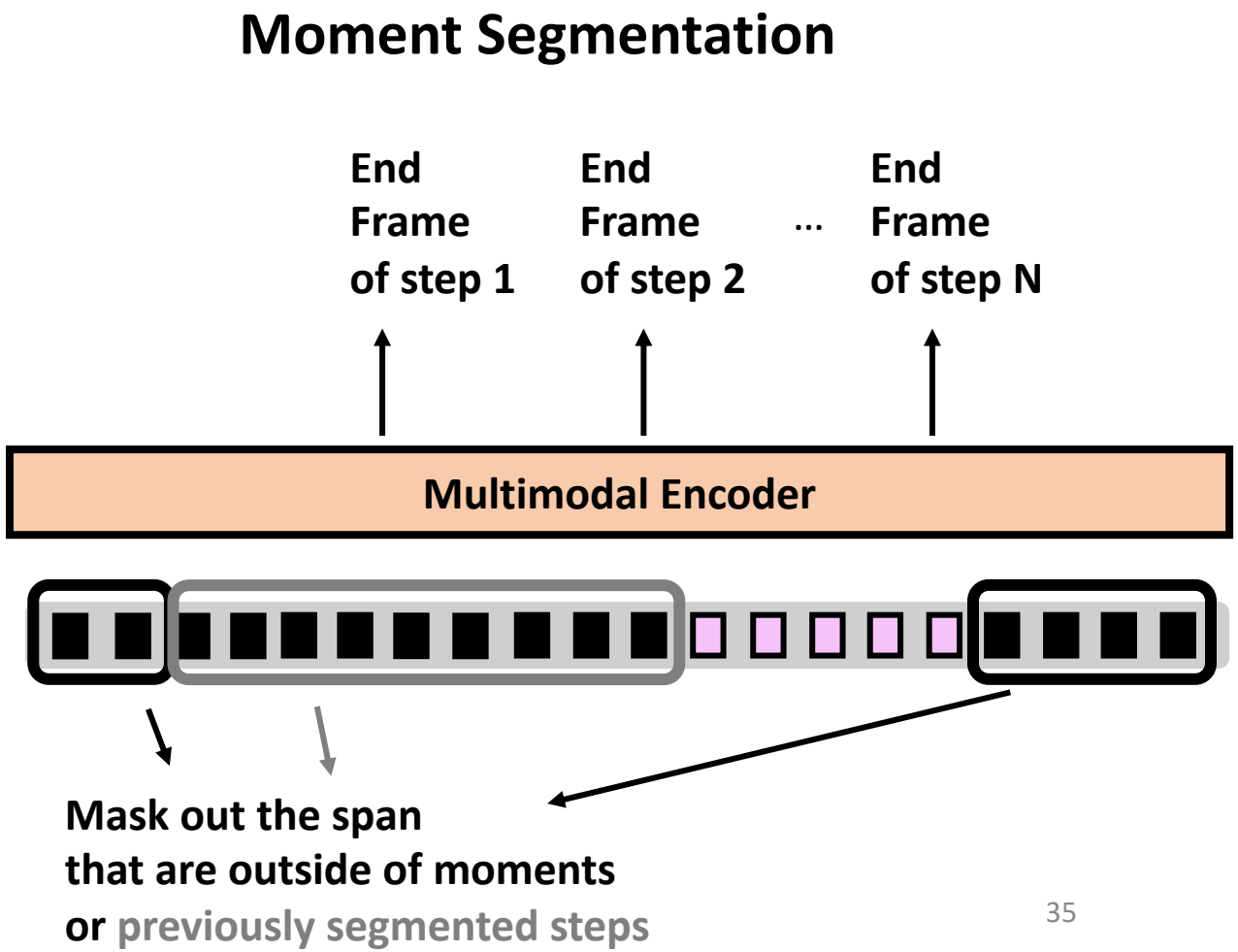
Find the right moment



Baseline Model for HiREST – Moment Segmentation

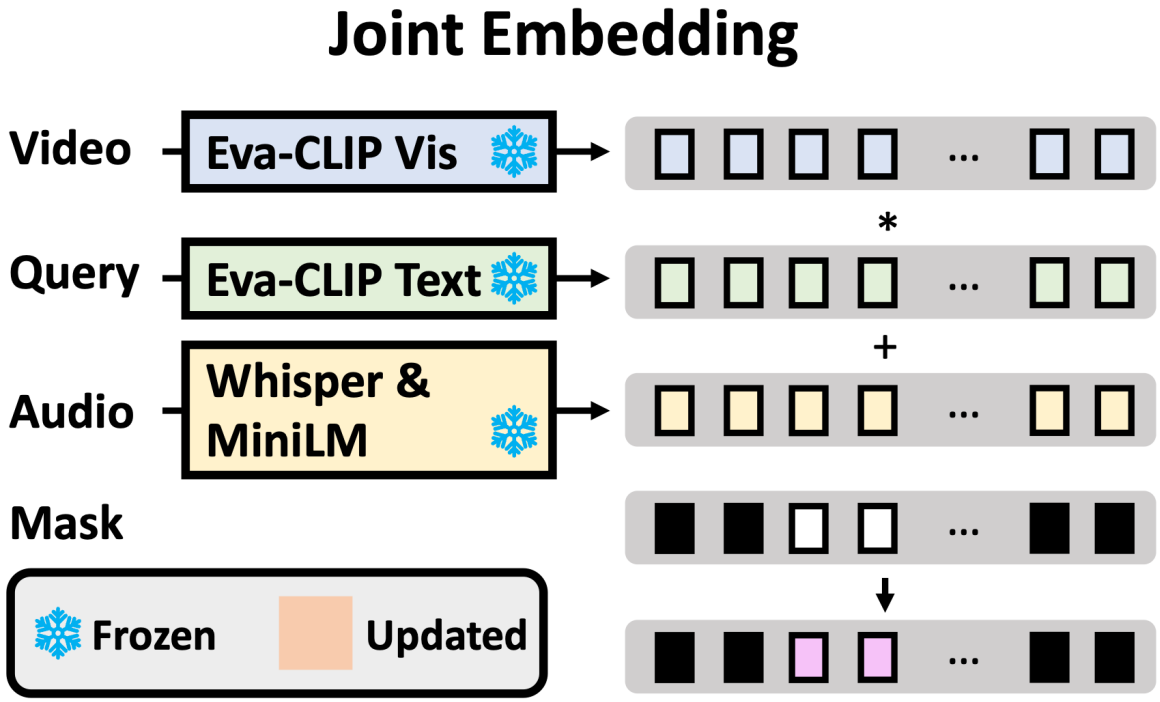
Summarize the moment in steps

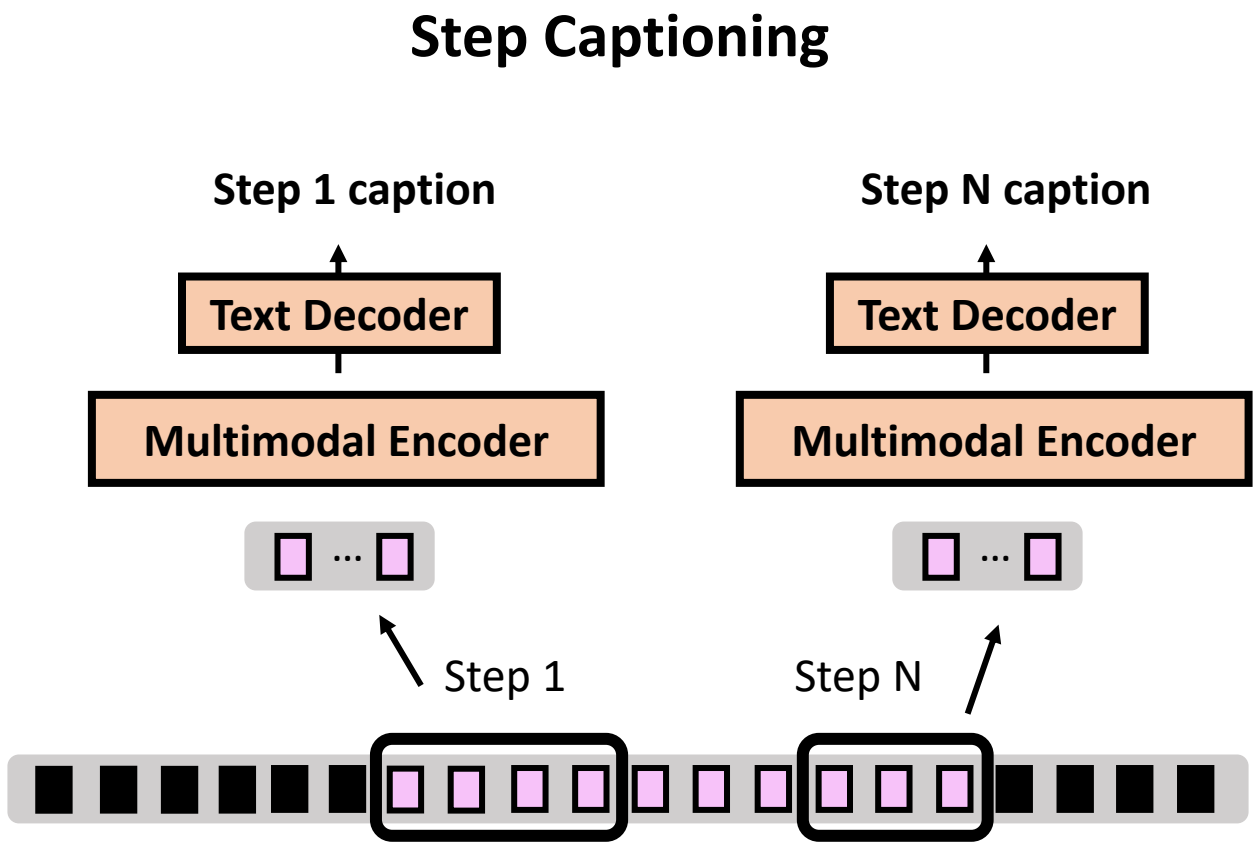




Baseline Model for HiREST – Step Captioning

Describe each step in text





Quantitative Evaluation

Joint end-to-end baseline model is similar to / better than task-specific models

Moment Retrieval

Model	FT	R@0.5	R@0.7
CLIP-B/32 (threshold=0.05)		21.01	9.02
CLIP-B/32 (8 frames left/right)		34.02	15.72
EVA-CLIP-G/14 (threshold=0.10)		19.33	7.86
EVA-CLIP-G/14 (8 frames left/right)		38.27	19.33
BMT		43.56	10.57
BMT	✓	71.91	**39.18**
Joint (Ours)	✓	**73.32**	32.60

Moment Segmentation

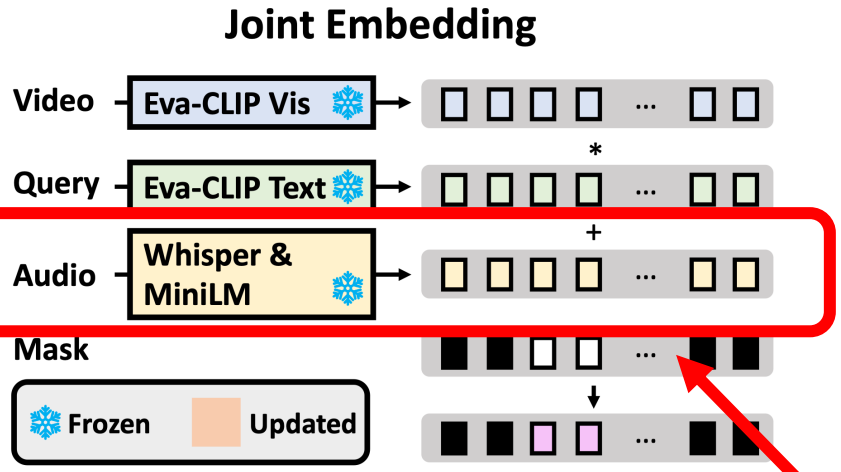
Model	FT	Recall@IoU		Precision@IoU	
		0.5	0.7	0.5	0.7
SSIM@0.75 (32 frames)		12.24	5.27	26.32	10.05
SSIM@0.85 (32 frames)		25.03	9.79	**37.38**	**13.80**
BMT (1fps)		8.24	3.71	20.95	7.96
BMT (1 fps)	✓	34.07	12.35	24.71	8.93
Joint (Ours) (1 fps)	✓	**37.50**	**14.76**	28.52	10.84

Step Captioning

Model	FT	METEOR	CIDEr	SPICE	Entail. (%)	BERT-S	CLIP-S
BMT		2.23	1.04	1.41	1.17	0.83	0.21
SwinBERT		5.12	13.31	4.65	5.86	0.85	**0.23**
BMT	✓	3.84	6.72	1.05	30.68	0.82	0.20
SwinBERT	✓	**5.94**	**24.66**	**6.67**	35.09	**0.86**	**0.23**
Joint (Ours)	✓	4.13	23.01	3.54	**43.88**	**0.86**	**0.23**

Quantitative Evaluation

Audio is useful

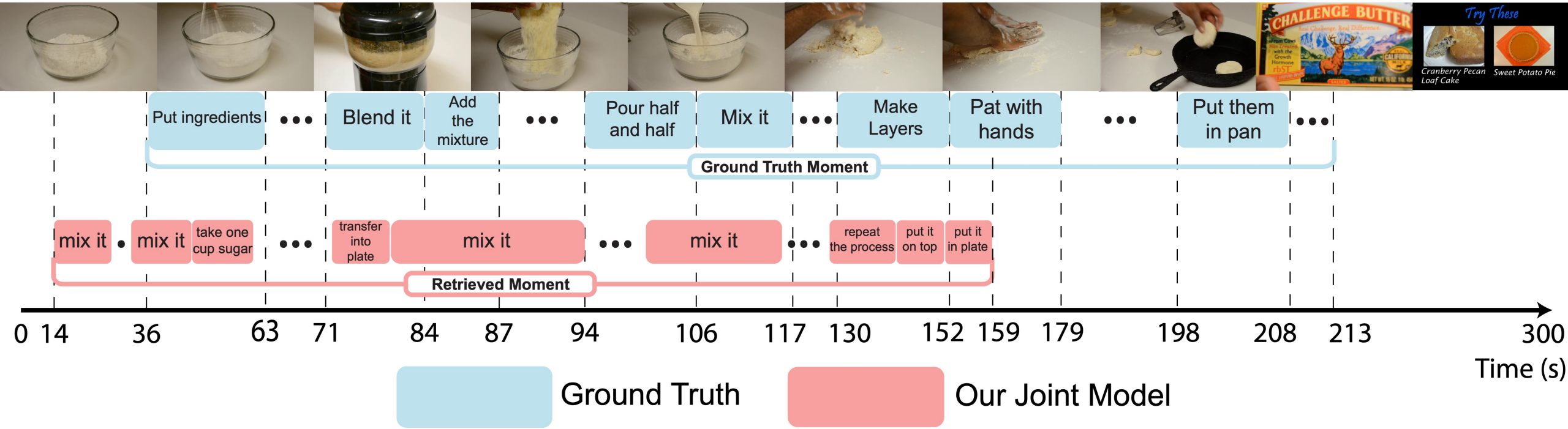


Model	FT	Moment Retrieval		Moment Segmentation		Step Captioning
		R@0.5	R@0.7	R@0.7	P@0.7	CIDEr
With Audio						
BMT	✓	71.9	39.2	12.4	8.9	6.7
Joint (Ours)	✓	73.3	32.6	14.8	10.8	23.0
Without Audio						
BMT	✓	62.6 (-9.3)	32.34 (-6.8)	10.4 (-2.0)	7.4 (-1.6)	6.1 (-0.6)
Joint (Ours)	✓	70.7 (-2.6)	20.6 (-12.0)	13.5 (-1.3)	10.0 (-0.8)	15.2 (-7.8)

Removing audio input hurts performance

End-To-End Hierarchical Video Information Retrieval

Given a text query ‘How to make butter biscuits’, our joint model
1) predicts a relevant moment from a video, 2) segments the moment into steps, and 3) describes the moment step-by-step.



Summary

1. HiREST Dataset

- A new benchmark that covers hierarchy in video search / multimodal summarization
- Provides high-quality / diverse step captions written by human annotators

2. Joint Baseline Model

- Provides starting point for all 4 tasks
- Shows comparable performance with task-specific baselines

Future work - Beyond 'How to XXX' queries

We hope HiREST fosters future work on end-to-end system for holistic multimodal retrieval / summarization

Data / Code / Demo

hirest-cvpr2023.github.io