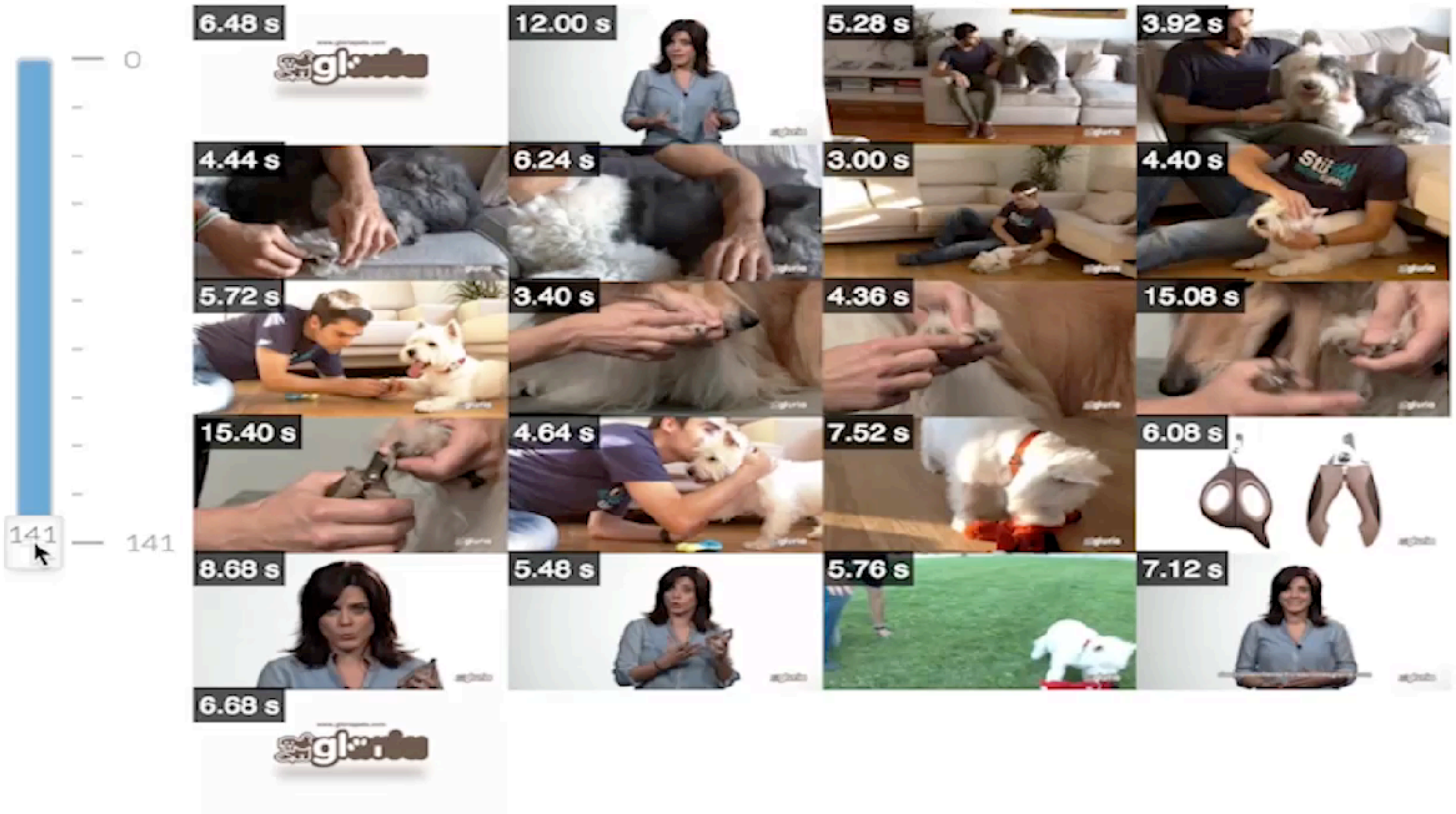


ElasticPlay

Interactive Video Summarization with Dynamic Time Budgets

Haojian Jin (CMU) Yale Song (Yahoo Research) Koji Yatani (UTokyo)

ElasticPlay

a javascript library that enables **interactive** video summarization
a new interface to **present/consume** video analysis in new ways.

0
141
141
6.48 s
12.00 s
5.28 s
3.92 s
4.44 s
6.24 s
3.00 s
4.40 s
5.72 s
3.40 s
4.36 s
15.08 s
15.40 s
4.64 s
7.52 s
6.08 s
8.68 s
5.48 s
5.76 s
7.12 s
6.68 s
gloria

Interactive Video Summarization => Human + Algorithms

introduction

video consumption

US adults spend 5.5 hours with video content (TV and online videos) per day.[1]

The average internet video length is 4.5 minutes. [2]

The average watch time of a single Internet video is 2.7 minutes. [3]

1. https://www.emarketer.com/Article/US-Adults-Spend-55-Hours-with-Video-Content-Each-Day/1012362
2. https://www.minimatters.com/youtube-best-video-length/
3. https://blog.kissmetrics.com/increase-youtube-video-engagement/

video consumption

US adults spend 5.5 hours with video content (TV and online videos) per day.[1]

The average internet video length is 4.5 minutes.[2]

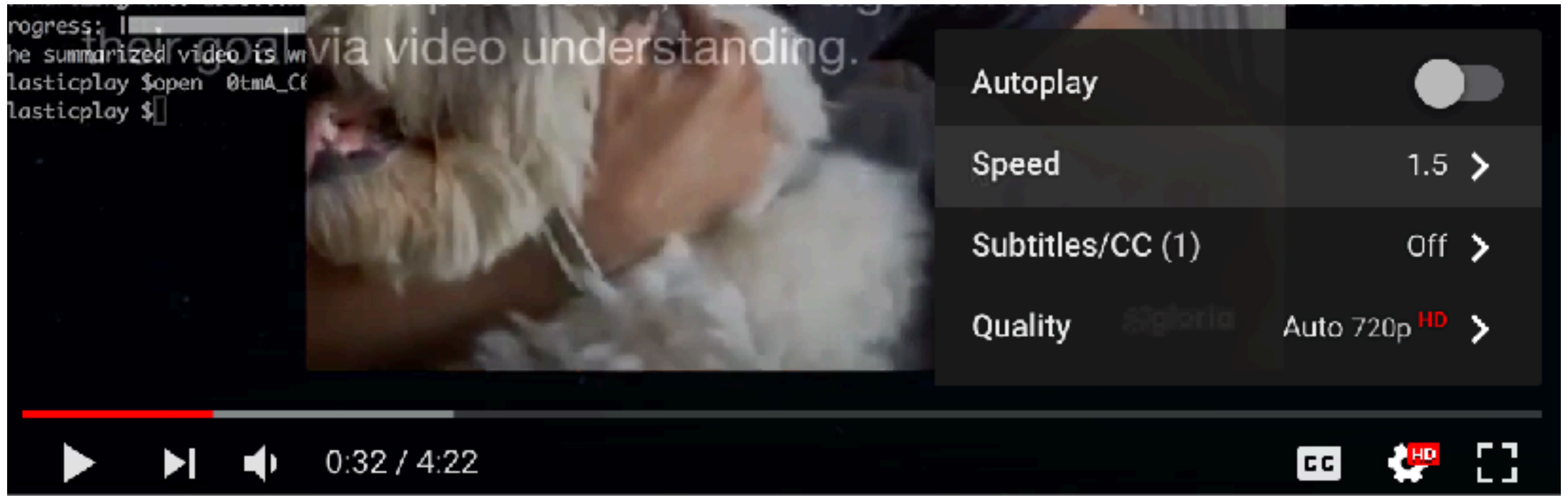
The average watch time of a single Internet video is 2.7 minutes.[3]

Users skipped 40% of the video content regularly.

1. https://www.emarketer.com/Article/US-Adults-Spend-55-Hours-with-Video-Content-Each-Day/1012362
2. https://www.minimatters.com/youtube-best-video-length/
3. https://blog.kissmetrics.com/increase-youtube-video-engagement/

video player interface

Youtube:



```
rogress: |
he summarized video is w
lasticplay $open  0tmA_C(
lasticplay $
```

their goal via video understanding.

Autoplay

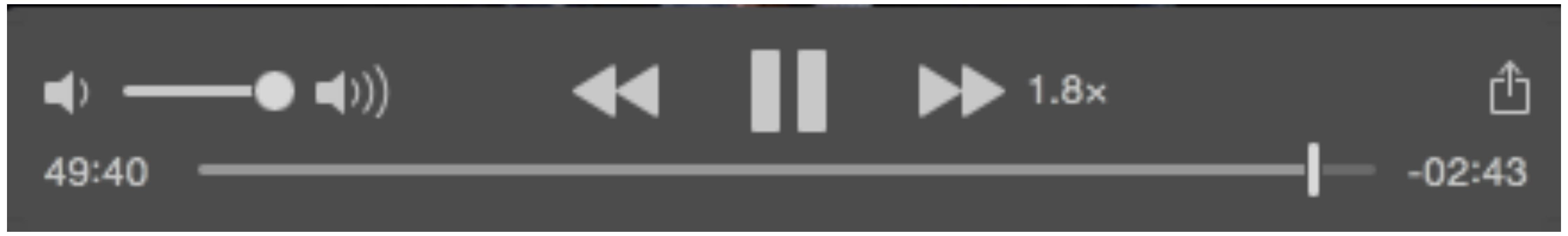
Speed 1.5

Subtitles/CC (1) Off

Quality Auto 720p HD

0:32 / 4:22

QuickPlay:



1.8×

49:40

-02:43

Timeline widget

Variable playback speed

Fast-forward

Timeline widget

Variable playback speed

Fast-forward

kodak projector assembly (1952)

a new kind of **user-centered** video interface?

a user expresses her **needs** through the interface,
the algorithms find a **global optimal** playback plan to fit that needs.
the use can then interact with the video by **updating** her context.

watch a 40-min video in a 30-min trip

static video summarization

automate the skipping process entirely

based on the desired length of a summary

static video summarization

automate the skipping process **entirely**

trial and error tuning

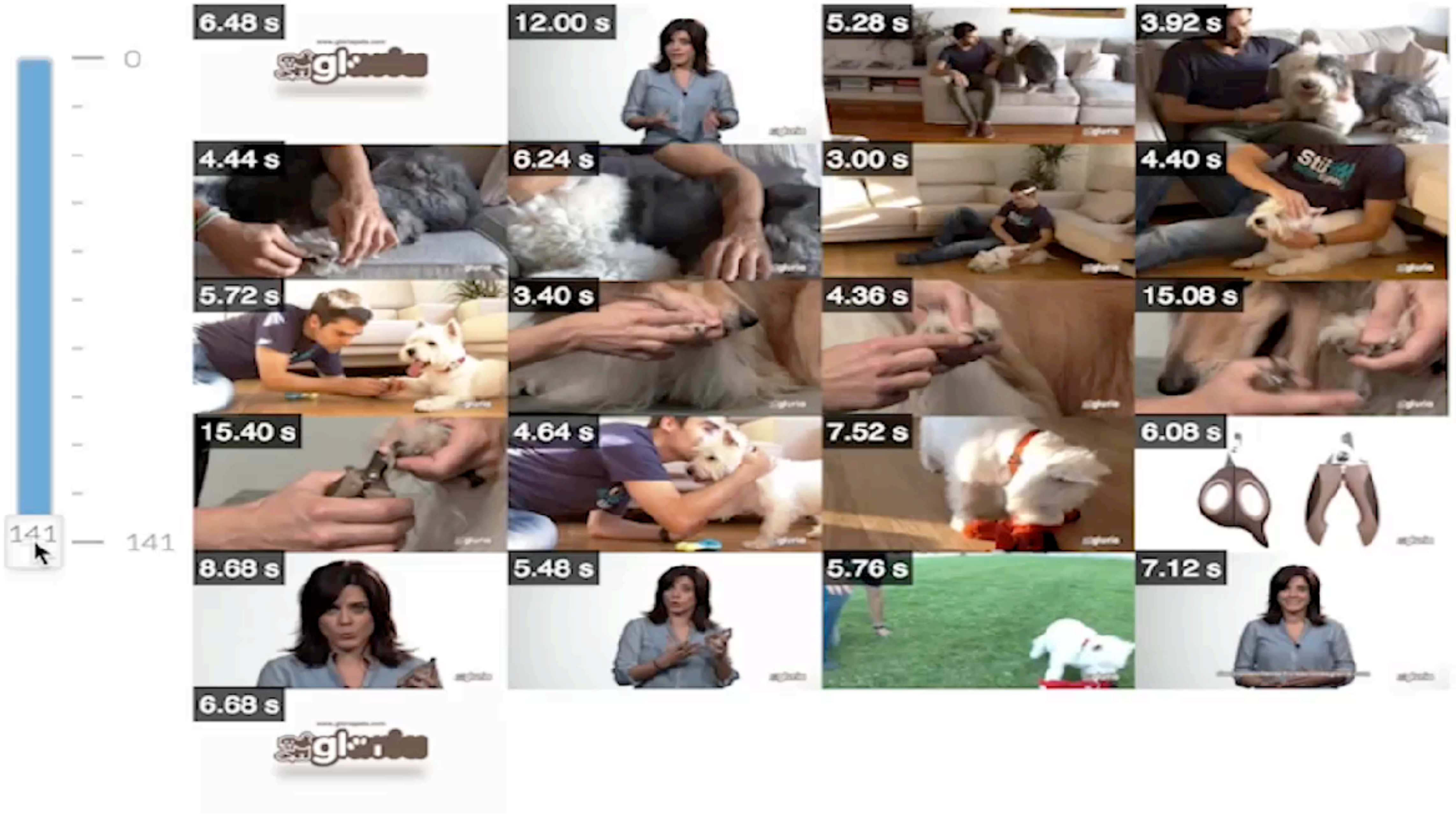
based on **the desired length** of a summary

context, personal preference, …

interactive
~~static~~ video summarization

real-time, transparent

users can live-tune the summarization on-the-fly until they are satisfied.



ElasticPlay

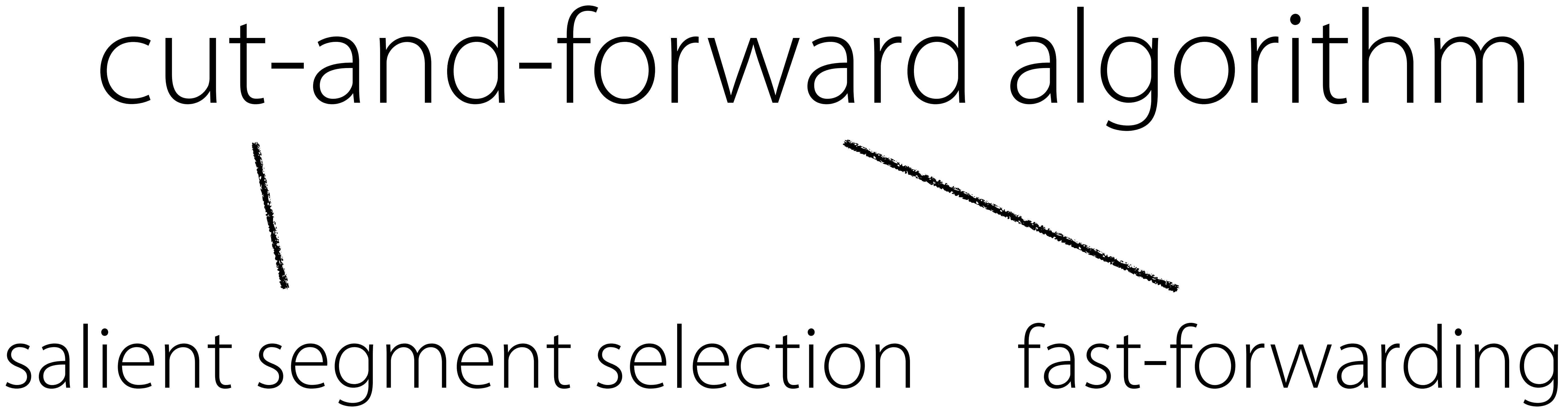
1) shortening strategy

2) exploration through interactivity

1 cut-and-forward algorithm

cut-and-forward algorithm

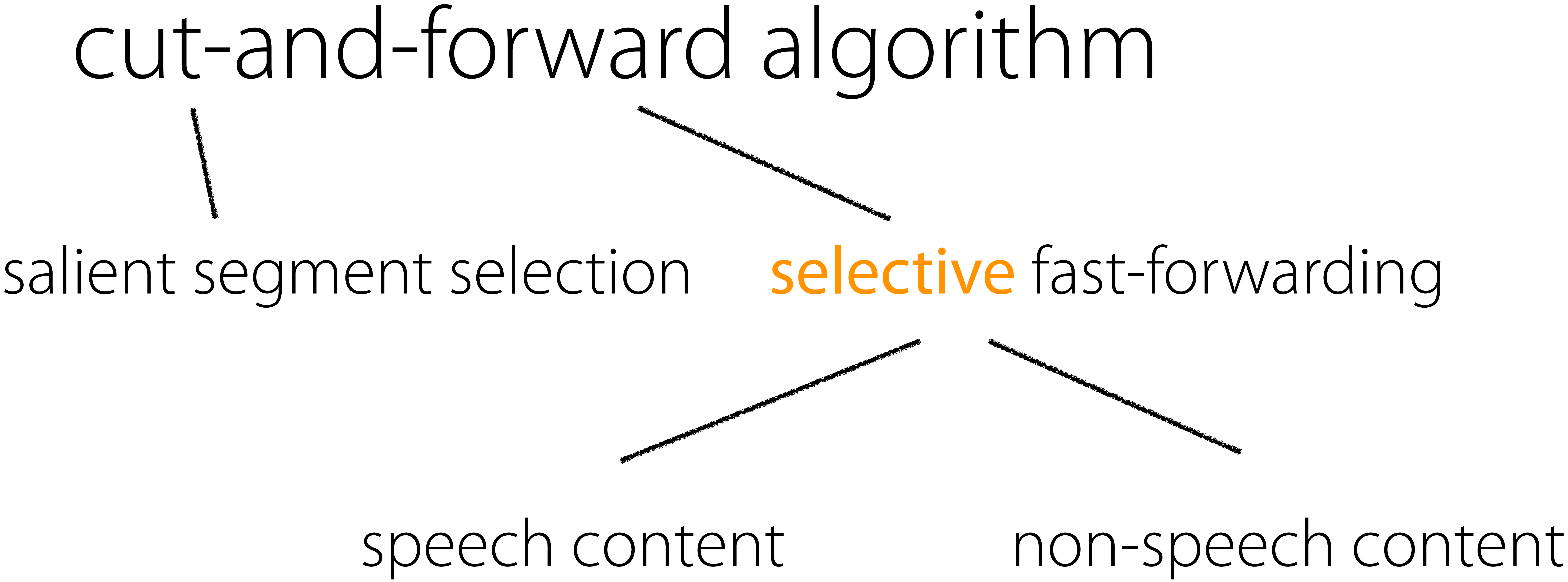
salient segment selection



cut-and-forward algorithm

salient segment selection

fast-forwarding



cut-and-forward algorithm

salient segment selection

selective fast-forwarding

speech content

non-speech content

cut-and-forward algorithm

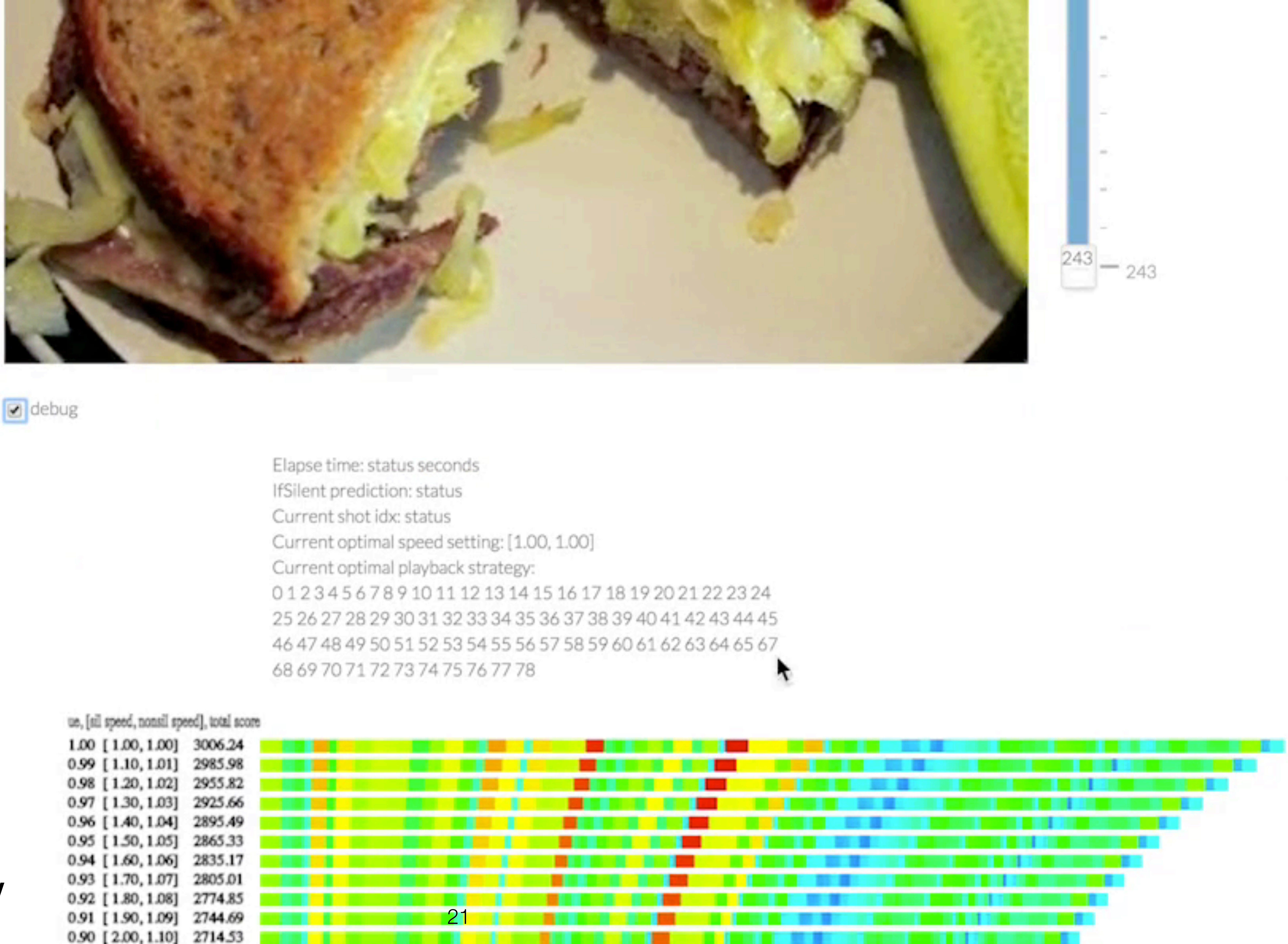
1. speed up the non-speech frames (most aggressive)

2. speed up the speech frames (moderately aggressive)

3. skip less interesting segments (less aggressive)

cut-and-forward algorithm

1.	speed up the non-speech frames (**most** aggressive)	6 s
2.	speed up the speech frames (moderately aggressive)	3 s
3.	skip less interesting segments (**less aggressive**)	1s
	saved total:	10s



☑ debug

Elapse time: status seconds
IfSilent prediction: status
Current shot idx: status
Current optimal speed setting: [1.00, 1.00]
Current optimal playback strategy:
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45
46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 67
68 69 70 71 72 73 74 75 76 77 78

```
ue, [sil speed, nonsil speed], total score
1.00 [ 1.00, 1.00]  3006.24
0.99 [ 1.10, 1.01]  2985.98
0.98 [ 1.20, 1.02]  2955.82
0.97 [ 1.30, 1.03]  2925.66
0.96 [ 1.40, 1.04]  2895.49
0.95 [ 1.50, 1.05]  2865.33
0.94 [ 1.60, 1.06]  2835.17
0.93 [ 1.70, 1.07]  2805.01
0.92 [ 1.80, 1.08]  2774.85
0.91 [ 1.90, 1.09]  2744.69
0.90 [ 2.00, 1.10]  2714.53
```

Debug view

comprehension model

$$AI_{all} = \sum_i p_i s_i$$

where p_i is the comprehension rate of the i-th shot,
and s_i is the importance score.

comprehension rate p_i

playback speed increases,

comprehension rate decreases.

it’s a linear relationship under certain thresholds.[1, 2]

[1] CinemaGazer: A System for Watching Videos at Very High Speed. AVI’12

[2] Adaptive fast playback-based video skimming using a compressed-domain visual complexity measure, ICME’04

selective fast-forwarding

the thresholds (*VT*) for speech and non-speech content are different.

the linear correlation factor factors (*k*) are different.[1, 2]

$$p = k \times (v - 1) + 1, \quad where \quad 1 \leq v \leq VT$$

[1] CinemaGazer: A System for Watching Videos at Very High Speed. AVI'12

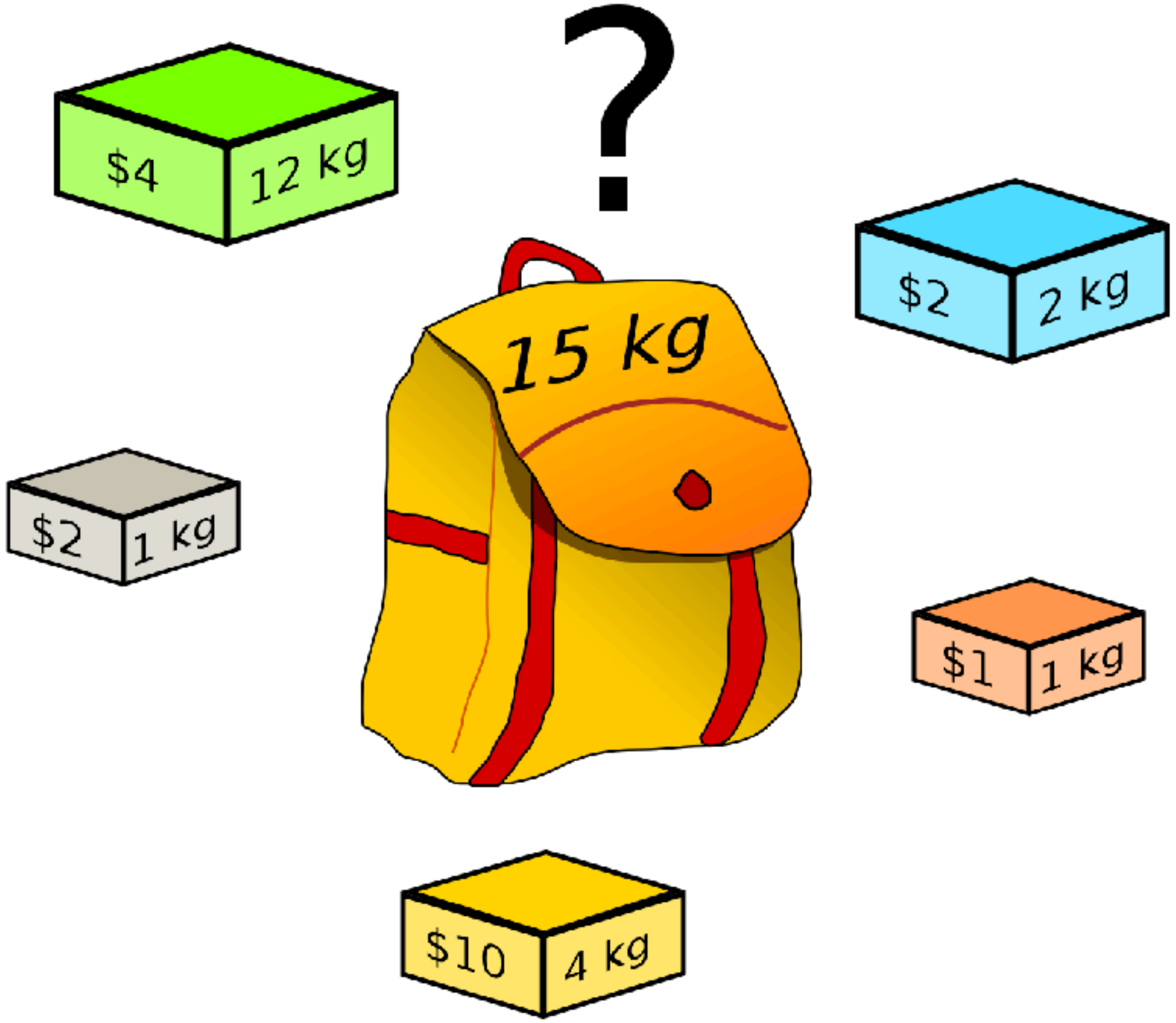
[2] Adaptive fast playback-based video skimming using a compressed-domain visual complexity measure, ICME' 04

mathematical optimization problem

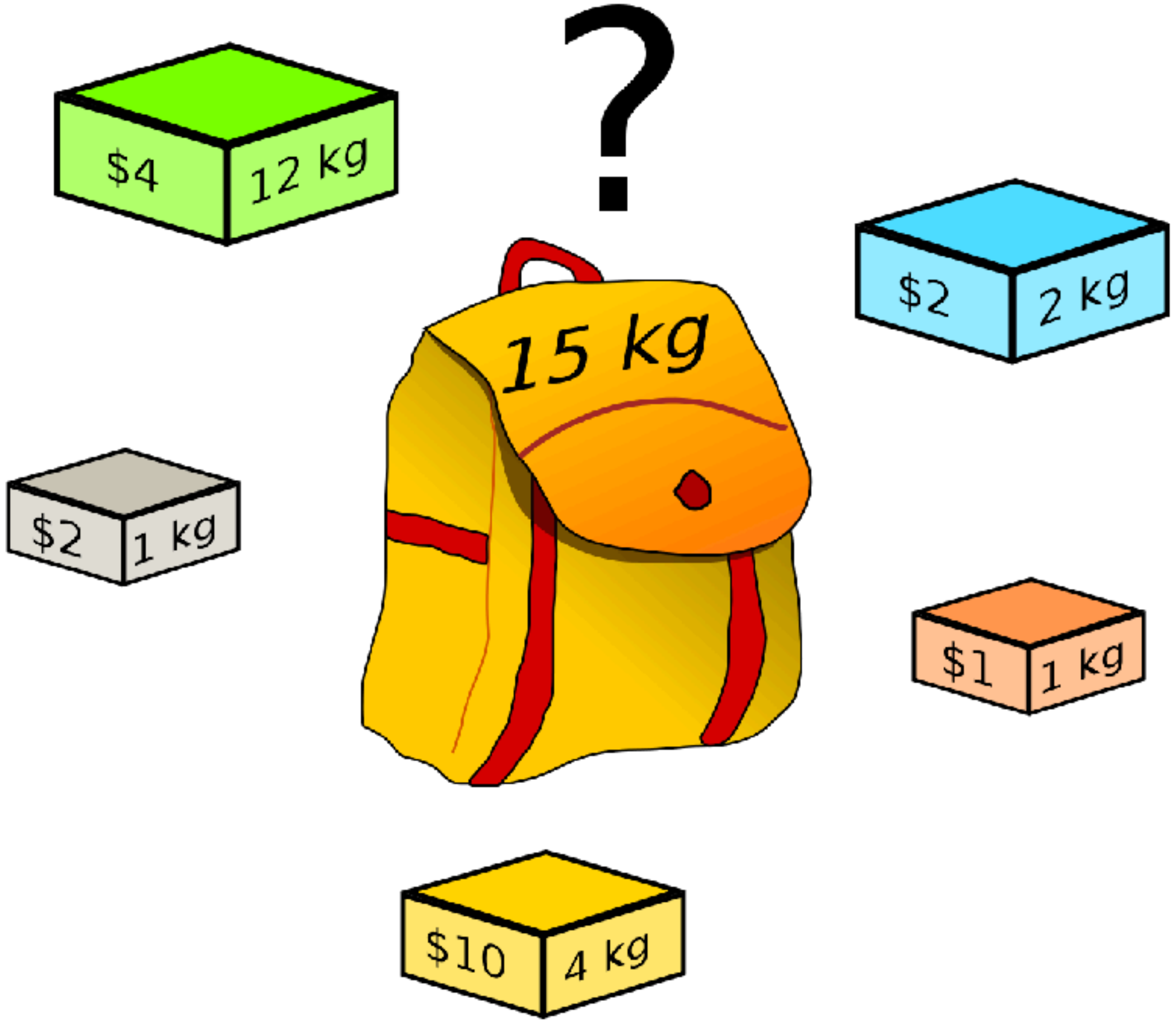
m-seconds video with *n* shots,
each shot has an importance score u_i ,

if we only have limited time,
which parts to skip or to fast-forward?

the 0/1 knapsack problem

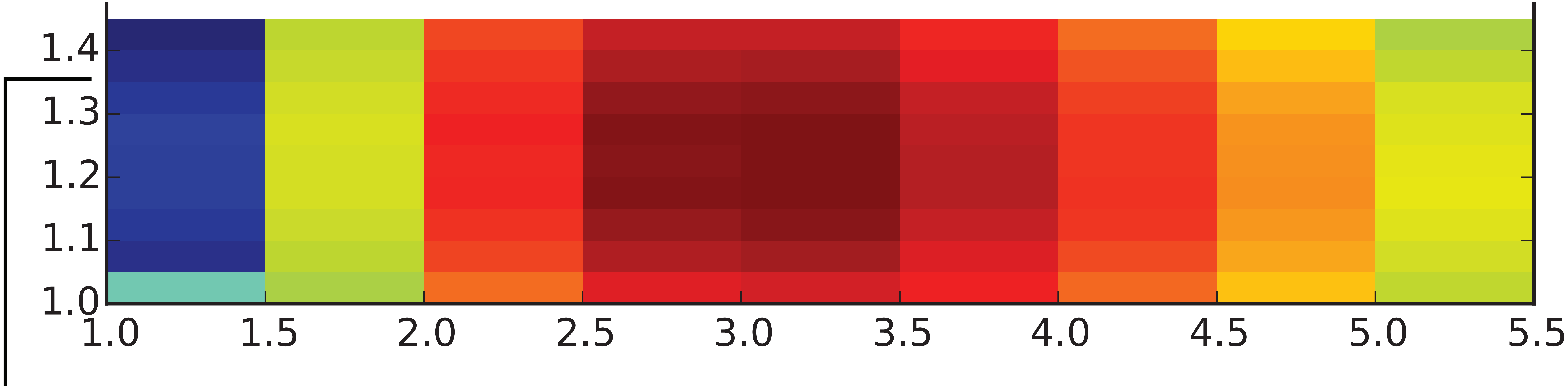


a variant of the 0/1 knapsack problem



alternative:
we can speed up a shot by losing some values.

solution search



y axis: speed for content with speech. **x axis**: speed for content without speech.
color: normalized score. Red indicates a higher score.

[3.0X, 1.15X] is the best solution.

2 exploration through interactivity

interactivity: realtime + transparency

provide immediate feedback for end users

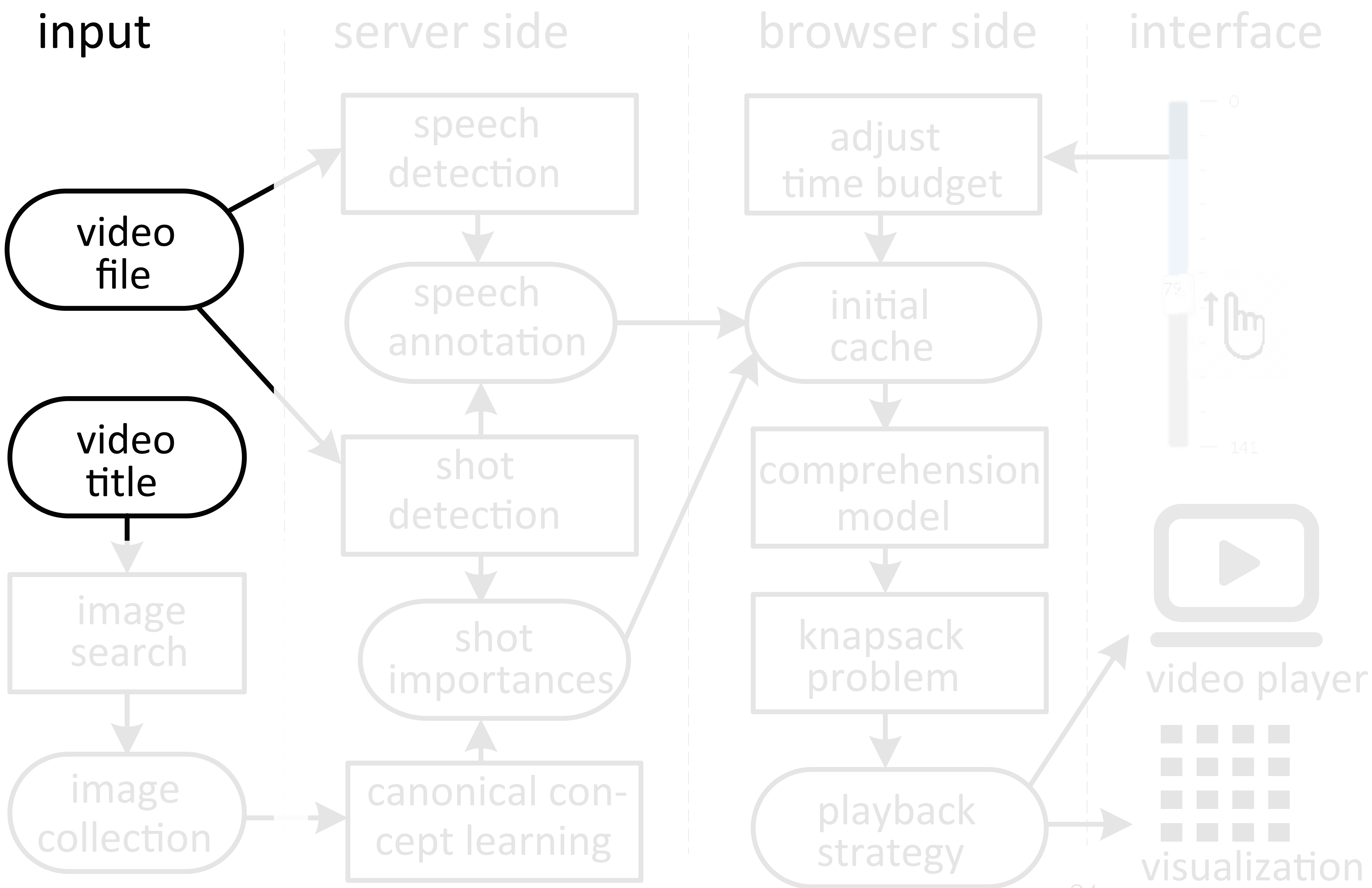
live-tuning

what-you-see-is-what-you-get[1]

provides a sense of the final output

[1] M. A. Hiltzik. Dealers of Lightning: Xerox PARC and the Dawn of the Computer Age. HarperBusiness, 2000.

system architecture overview



input

server side

browser side

interface

video file

video title

image search

image collection

speech detection

speech annotation

shot detection

shot importances

canonical concept learning

adjust time budget

initial cache

comprehension model

knapsack problem

playback strategy

video player

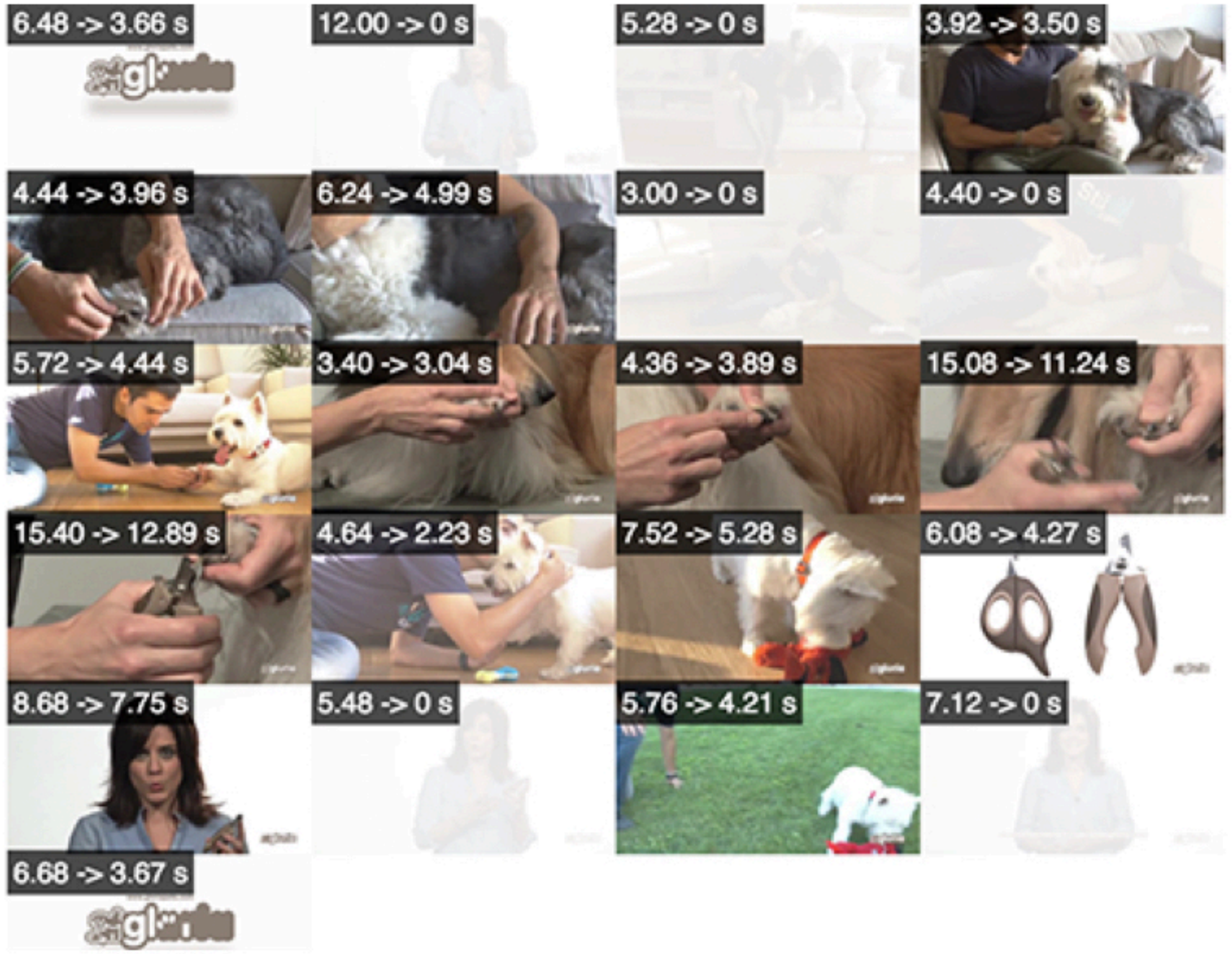
visualization

offline pre-processing

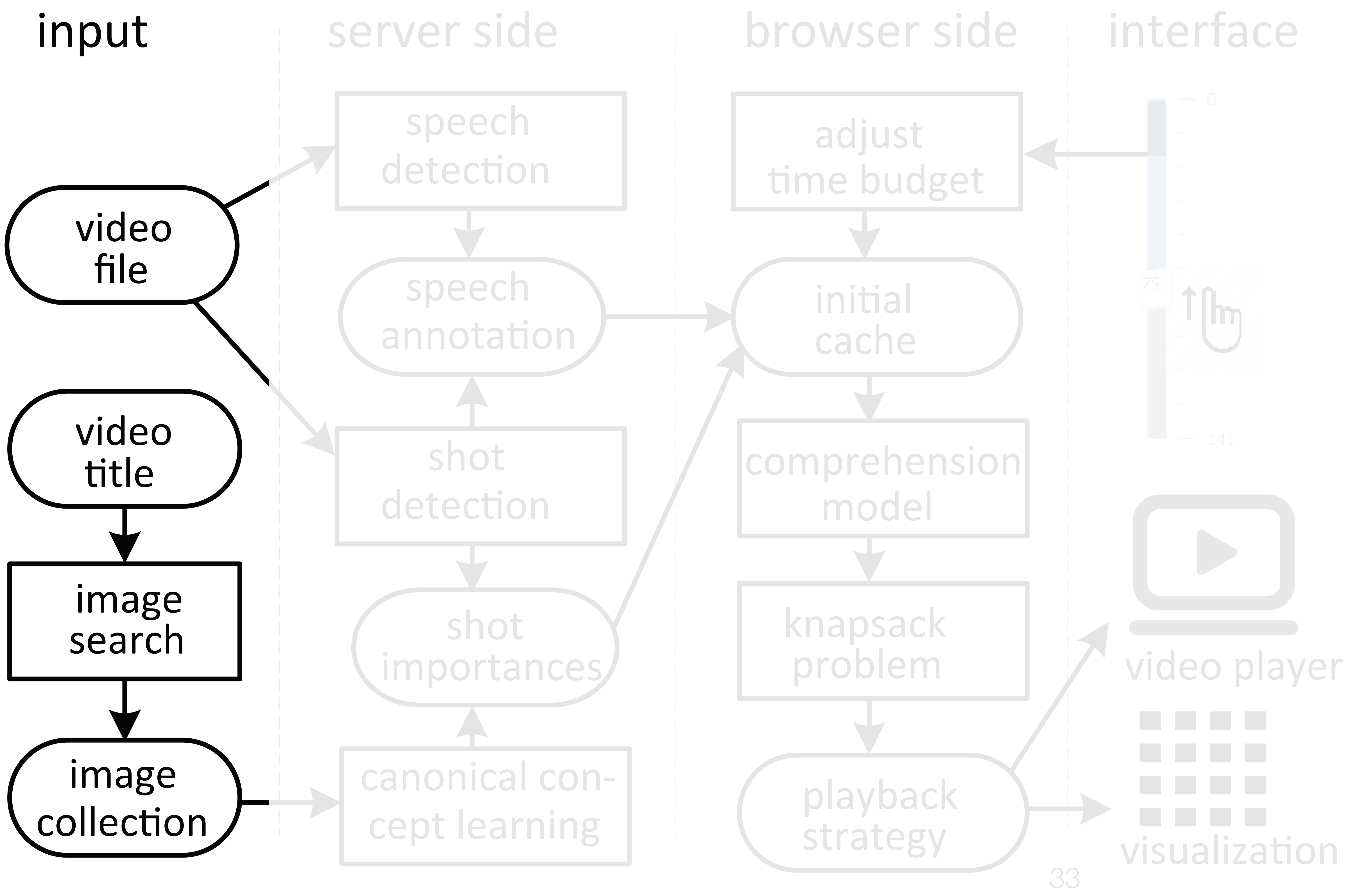
1. video segmentation

2. title-based importance score inference

3. speech detection



system architecture overview



solution search at interactive rate

initialization computing

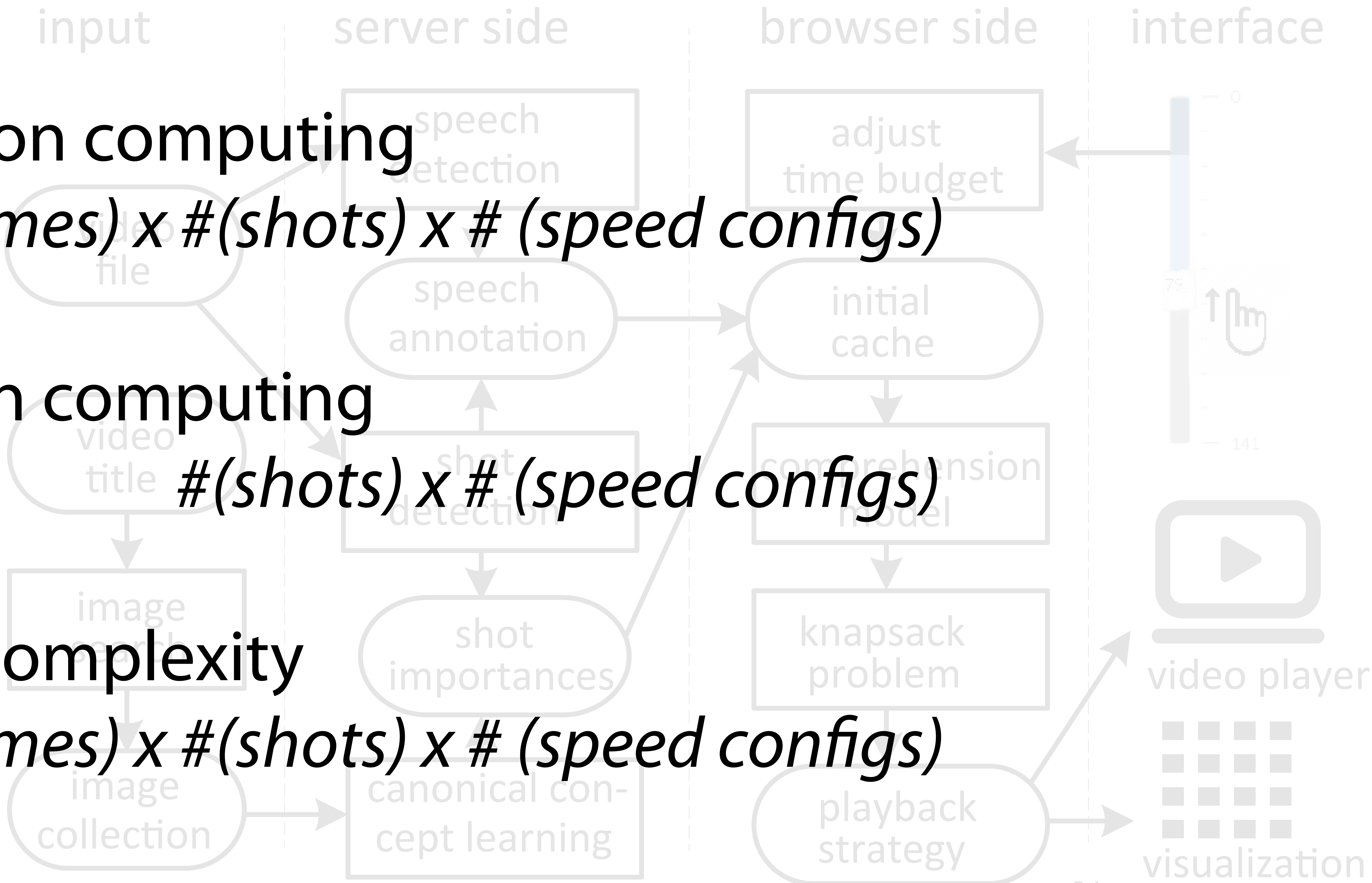
= *#(frames) x #(shots) x # (speed configs)*

interaction computing

= *#(shots) x # (speed configs)*

memory complexity

= *#(frames) x #(shots) x # (speed configs)*



dynamic programming for solution search

5-min video example, 80 shots, 9000 frames, 10 speed config.

initialization = 9000*80*10/2.4 gHz = 0.003 second
interaction = 800/2.4 gHz = 3e-7 second
cache size = 9000*80*10 bytes = 7.2 MB

scalable for videos up to 240 minutes for real-time processing.

on-the-fly interactive summarization?

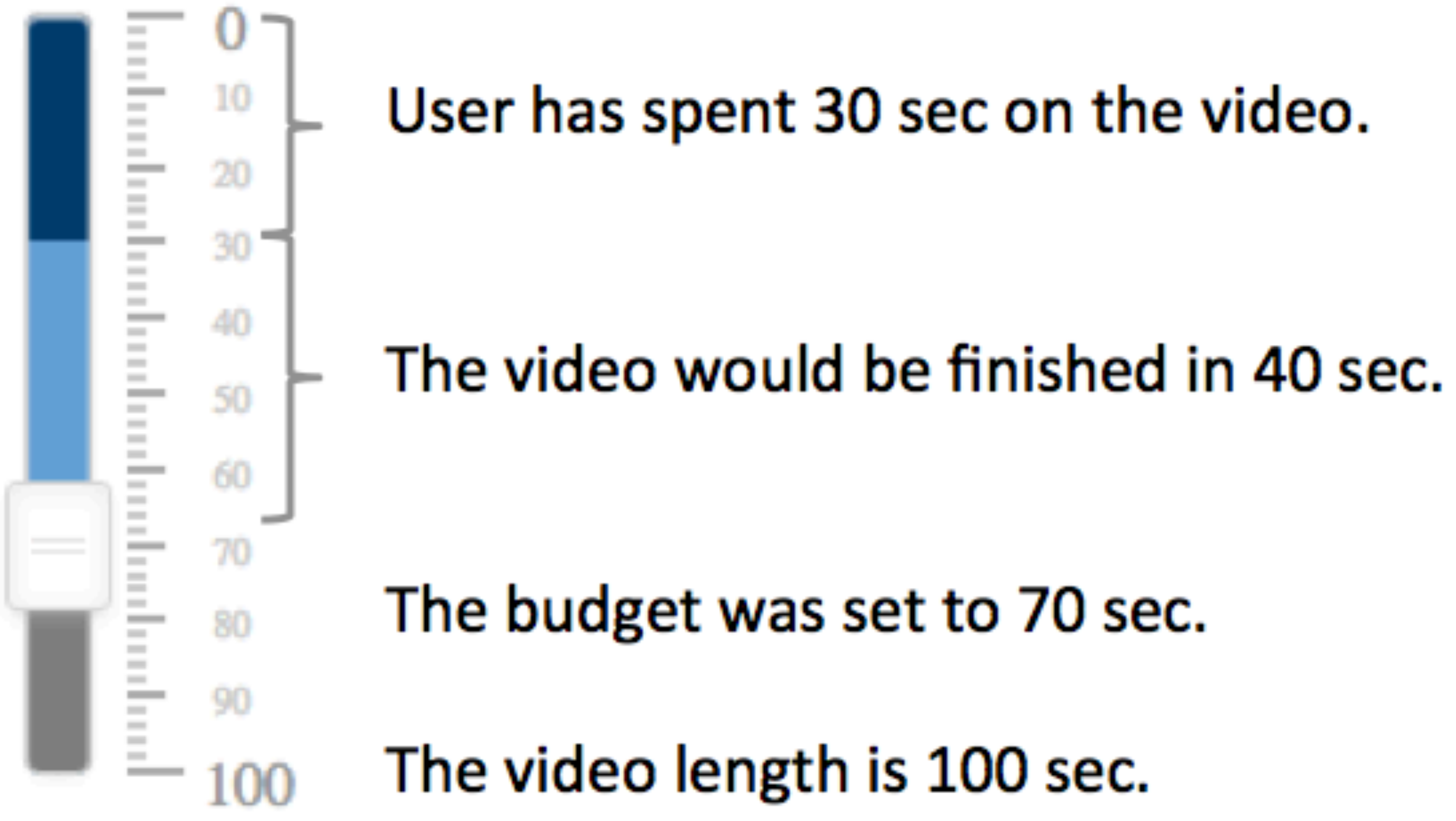
gloria
-1:31
0
110
141
6.48 -> 0 s
12.00 -> 0 s
5.28 -> 0 s
3.92 -> 0 s
4.44 -> 0 s
6.24 -> 5.37 s
3.00 -> 2.80 s
4.40 -> 4.11 s
5.72 -> 4.82 s
3.40 -> 3.18 s
4.36 -> 4.07 s
15.08 -> 12.33 s
15.40 -> 13.71 s
4.64 -> 2.83 s
7.52 -> 5.89 s
6.08 -> 4.77 s
8.68 -> 8.11 s
5.48 -> 5.12 s
5.76 -> 4.65 s
7.12 -> 6.28 s
6.68 -> 4.43 s

infinite # of summarization contexts

x seconds consumed *content,*

y seconds remaining *content,*

z seconds time budget.



User has spent 30 sec on the video.

The video would be finished in 40 sec.

The budget was set to 70 sec.

The video length is 100 sec.

reusable knapsack cache for interactivity

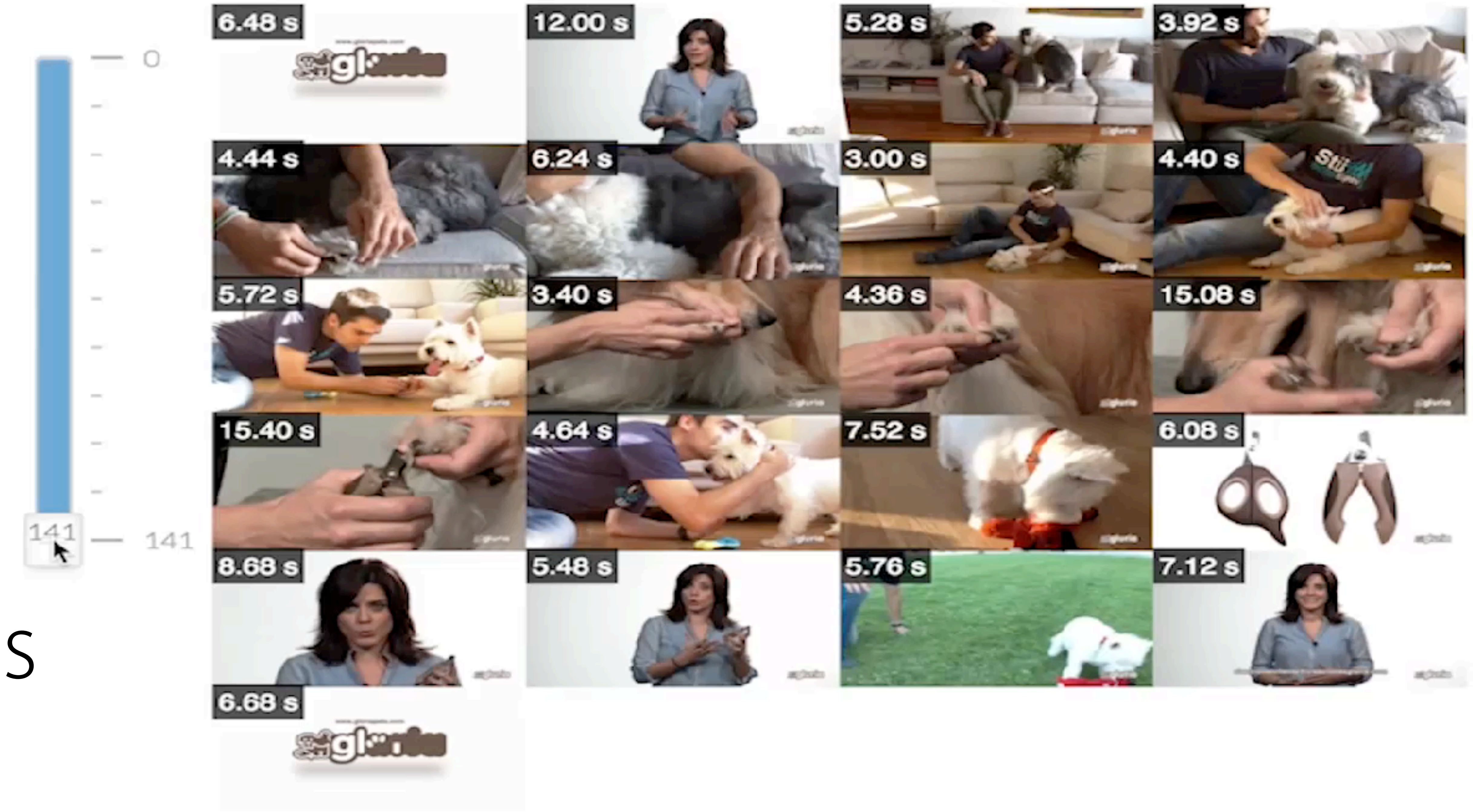
we always watch the video from the beginning
to the end.

designed a reversed cache design to reuse the
computing results. (details in the paper)

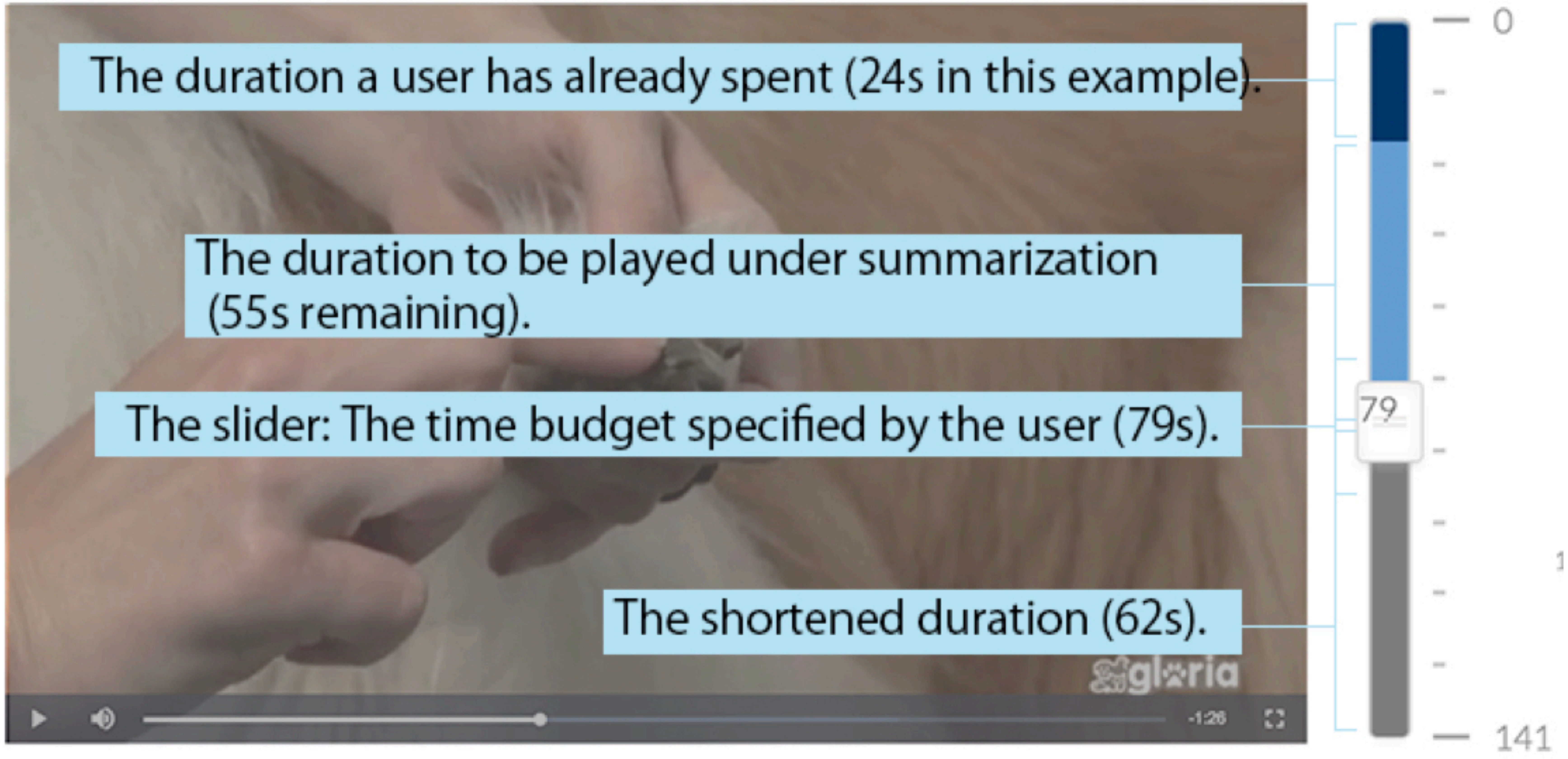
what-you-see-is-what-you-get

transparency -> compression rate

left-top number -> time differences



interactive summarization slider



The duration a user has already spent (24s in this example).

The duration to be played under summarization (55s remaining).

The slider: The time budget specified by the user (79s).

The shortened duration (62s).

evaluation

evaluation

1 quantitative algorithm

2 user experience of CaF-generated videos

3 ElasticPlay as a system

1 quantitative algorithm evaluations

content coverage

data set: TVSum, 50 videos
each video contains ratings by 20 people for every 5 seconds

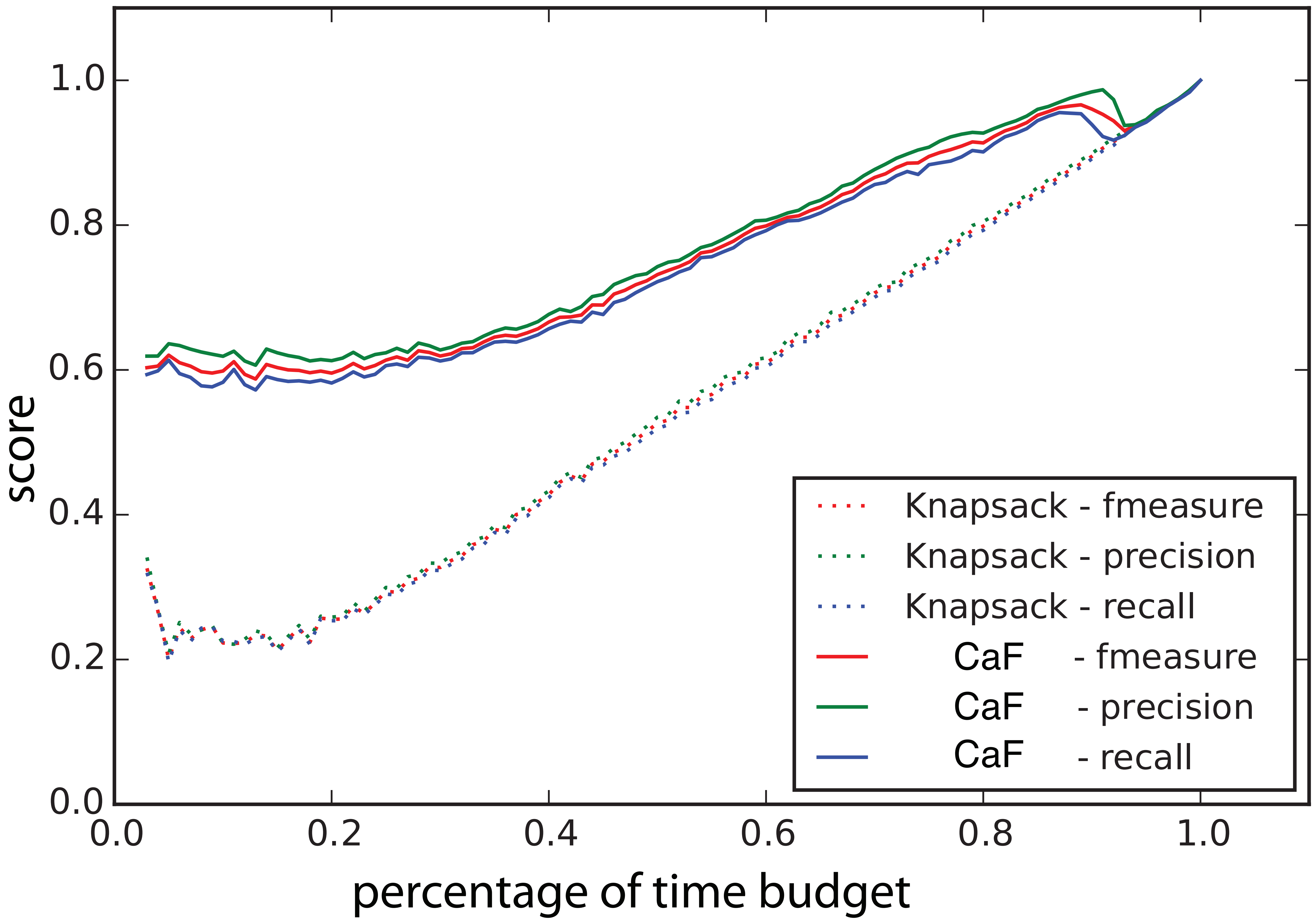
test: salient segment skipping [TVSum, CVPR’15]
cut-and-forward (hybrid approach)

metrics: F-1 score, Accuracy, Recall

content coverage

better relevance (recall)

higher quality (precision)



ⓐ F -measure, Precision, Recall

expected content comprehension

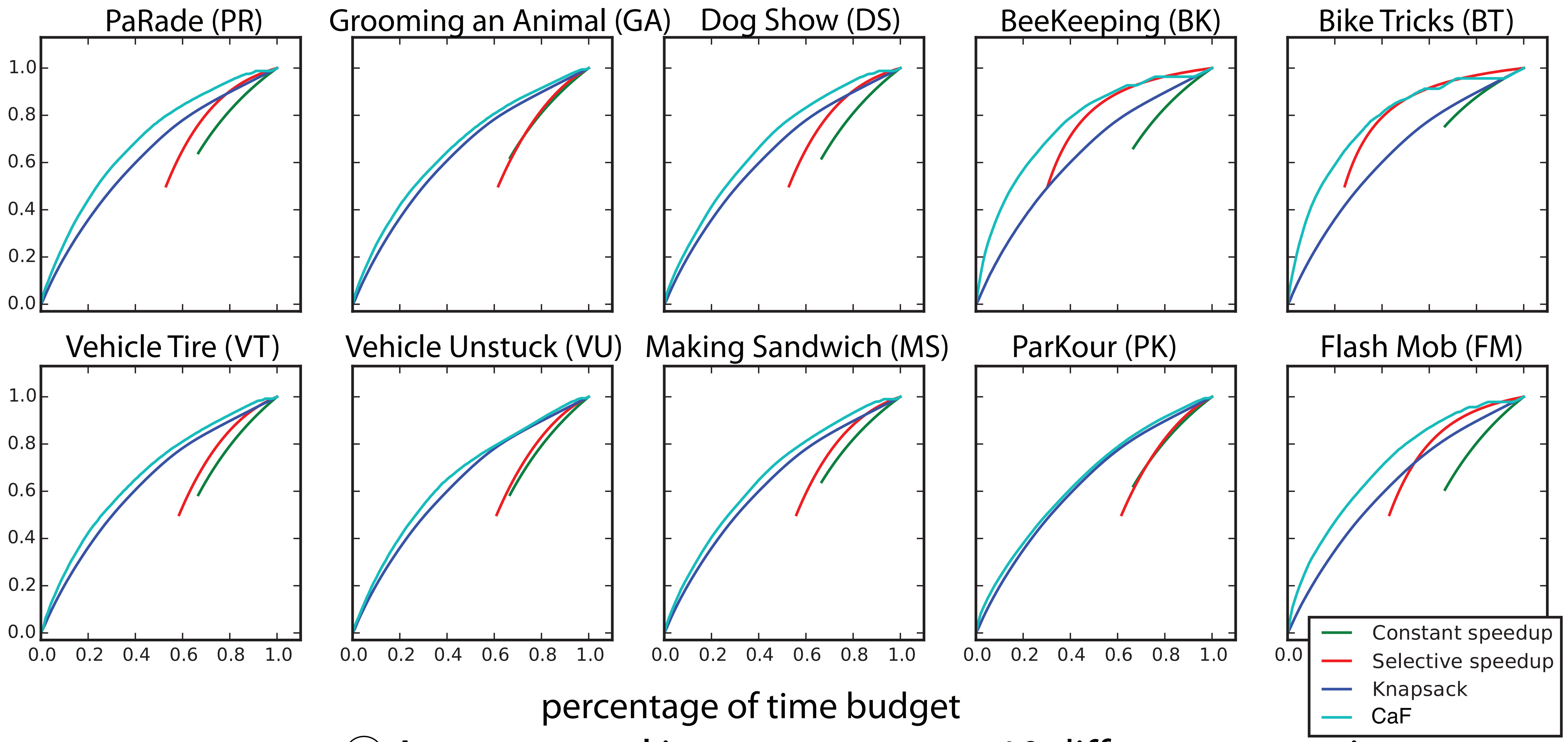
data set: TVSum, 50 videos
each video contains ratings by 20 people for every 5 seconds

test: salient segment skipping [TVSum, CVPR’15]
constant/selective fast-forwarding
cut-and-forward (hybrid approach)

metrics: comprehension model output score

expected content comprehension

better comprehension across categories



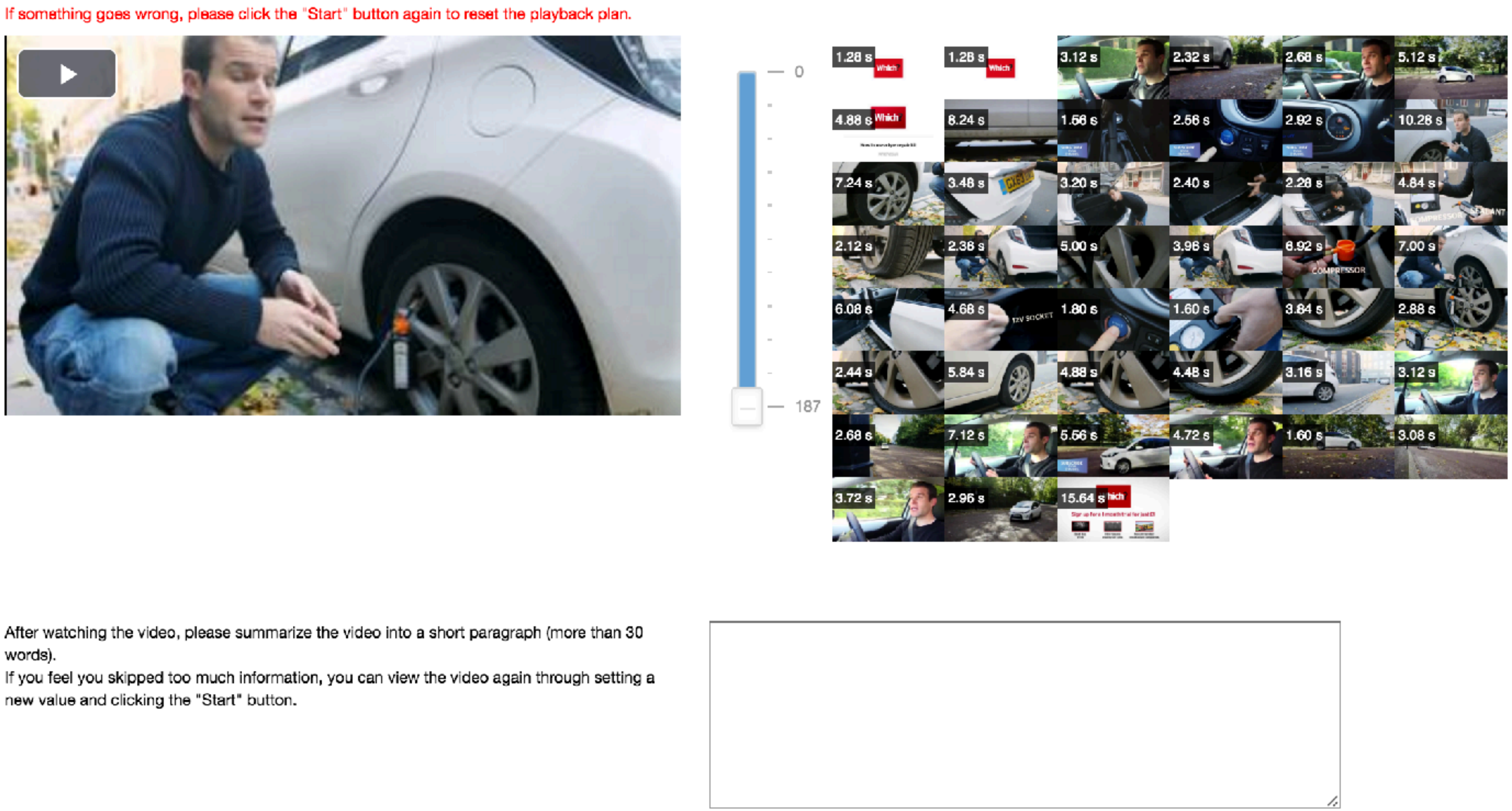
ⓑ Apparent total importance across 10 different categories

2 user experience of CaF-generated videos

=> paper

3 ElasticPlay as a system

user study interface



If something goes wrong, please click the "Start" button again to reset the playback plan.

After watching the video, please summarize the video into a short paragraph (more than 30 words).
If you feel you skipped too much information, you can view the video again through setting a new value and clicking the "Start" button.

study design

step 1: tutorial + one warmup task

step 2: four tasks in a randomized order

for each task, watch a video and write a summary

step 3: post-study survey

We record all the user behavior on the website.

participants stats

10 lab-participants

4 male, mean age 22.9, max=25, min=21

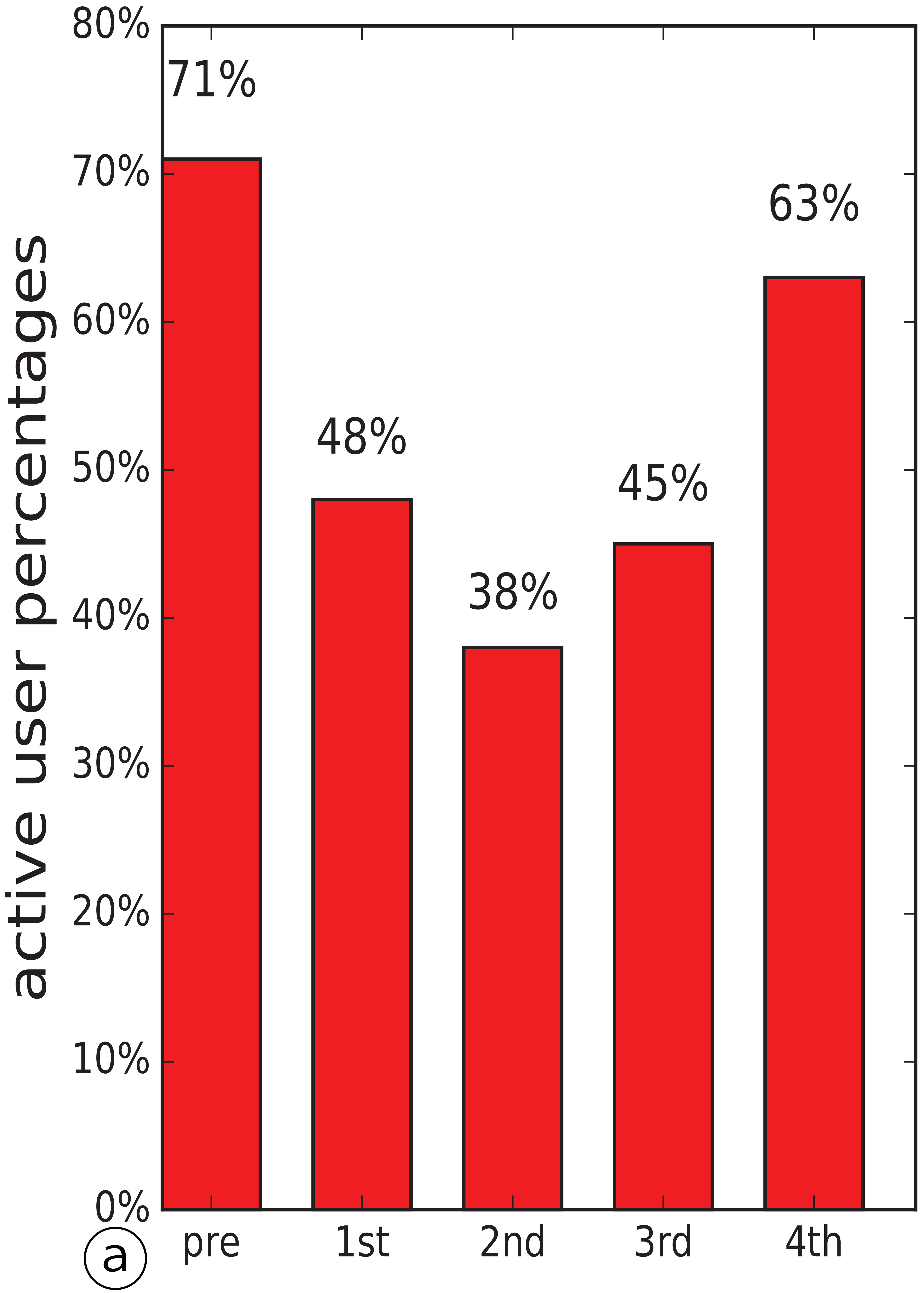
60 Amazon Mechanical Turk participants

study avg length: μ = 16.31 mins

summary avg length: μ = 57.38 words

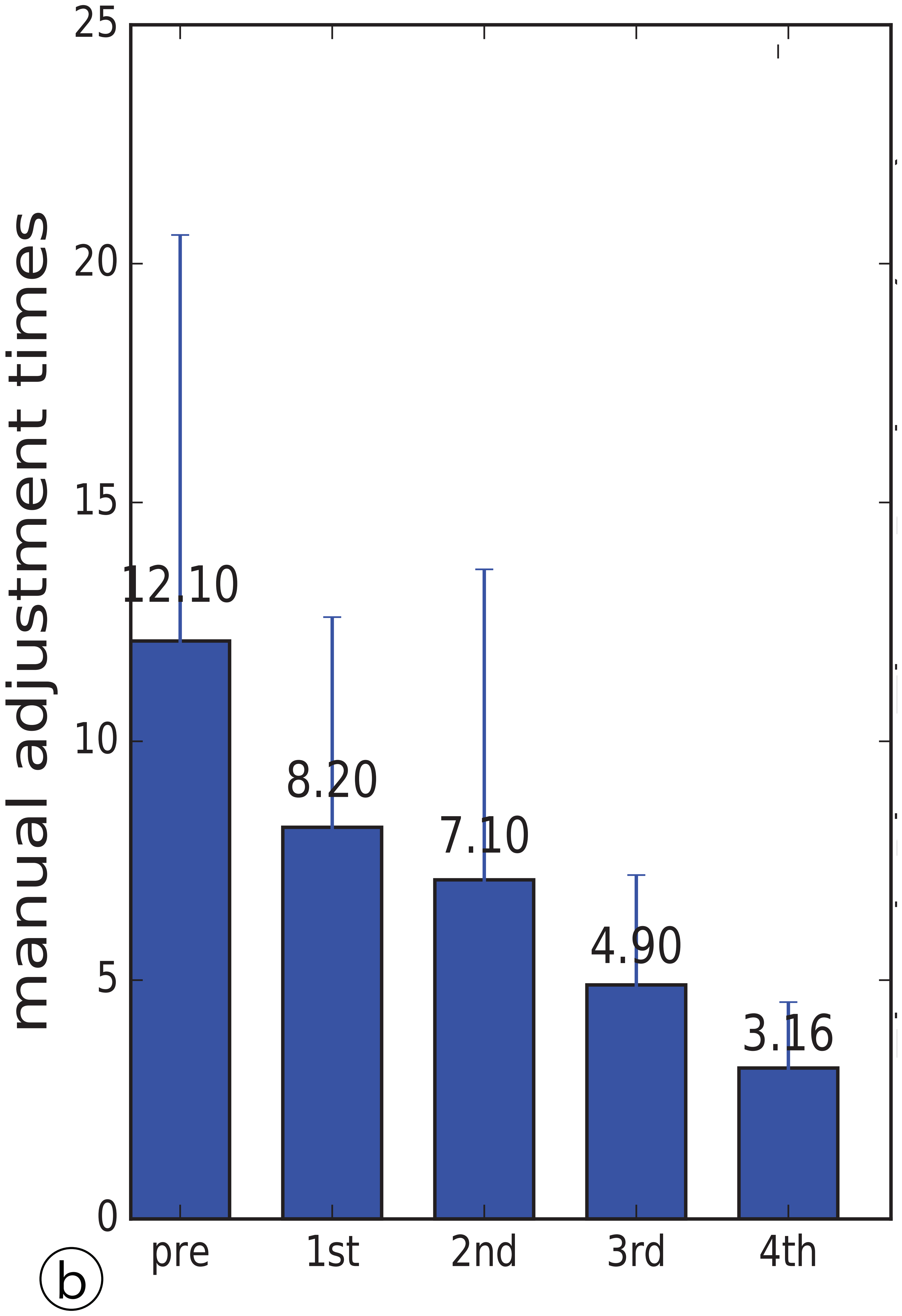
slider usages

the **consistent** usages shows participants are willing to **keep using** ElasticPlay.



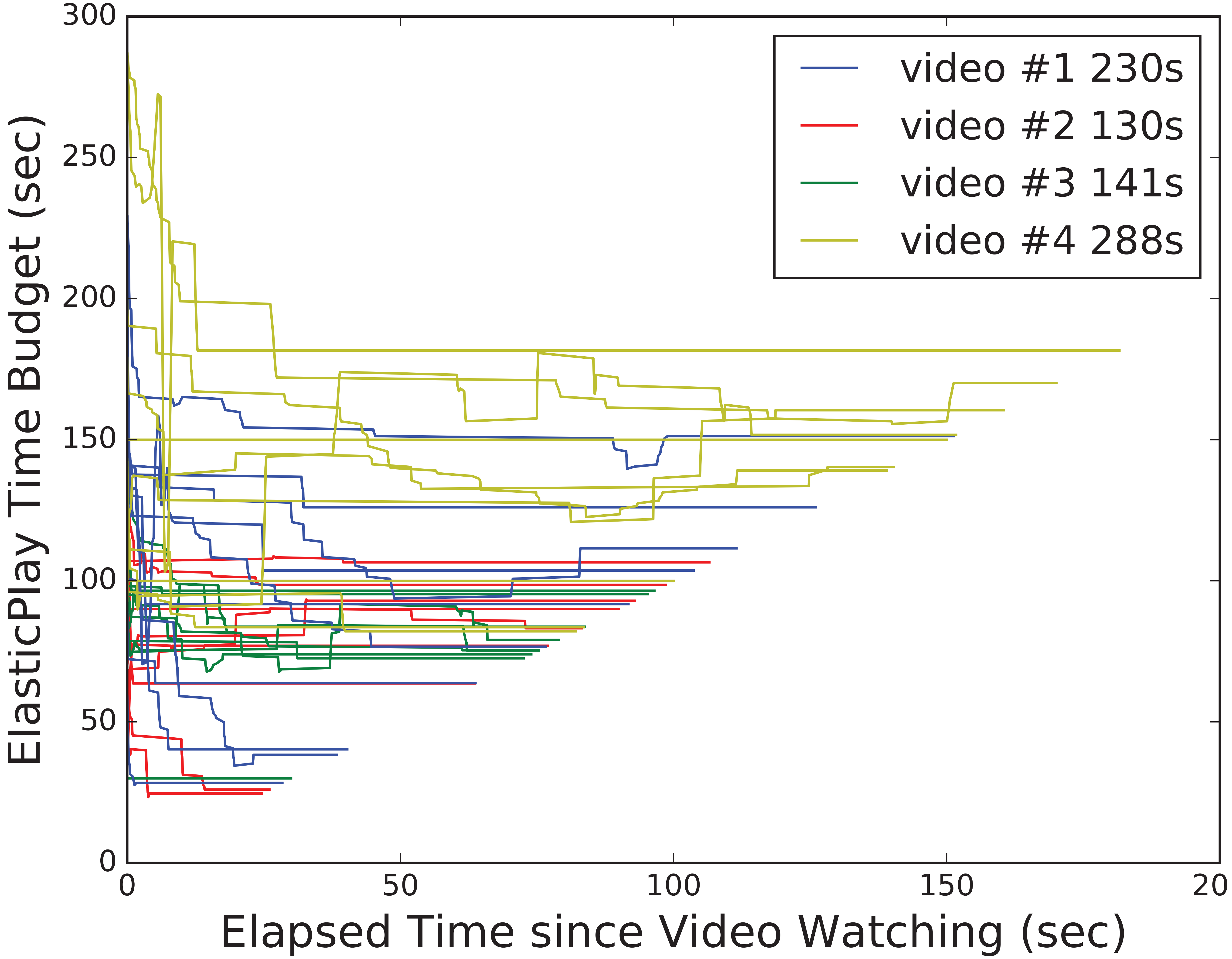
manual adjustment times

the **decreasing** trend suggests that participants were able to develop a correct **mental model** of ElasticPlay.



time budgets traits

most participants tended to **conservatively** estimate time budgets and gradually tuned them **during watching**.



ⓒ Time Bugest Adjustment Trace along Elaspse Time

conclusion

interactive video summarization
through dynamic time budget

the Cut-and-Forward algorithm that
combines salient segment selection and selective fast-forwarding

our evaluations suggest the benefits of
increased transparency and interactivity.

ElasticPlay

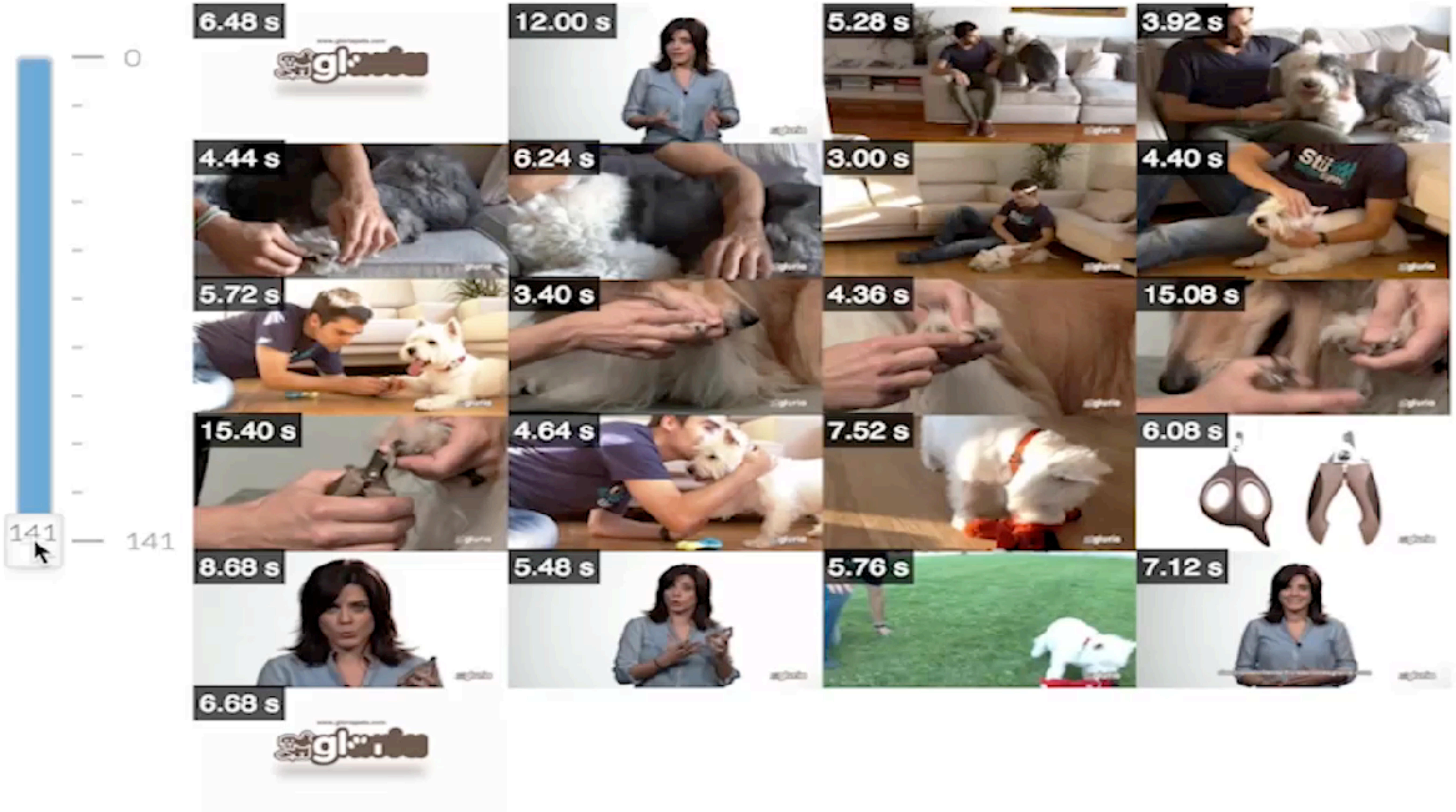
Interactive Video Summarization

Human + Algorithms

users have direct control over the summarization procedure,

algorithms help users achieve their goal via video understanding.

Live demo at:
bit.ly/elasticplay



ElasticPlay

Interactive Video Summarization with Dynamic Time Budgets

Haojian Jin (CMU) Yale Song (Yahoo Research) Koji Yatani (UTokyo)

Acknowledgments: Jason Hong