Lecture 11:

The Present and Future of Video Conferencing Systems

Visual Computing Systems

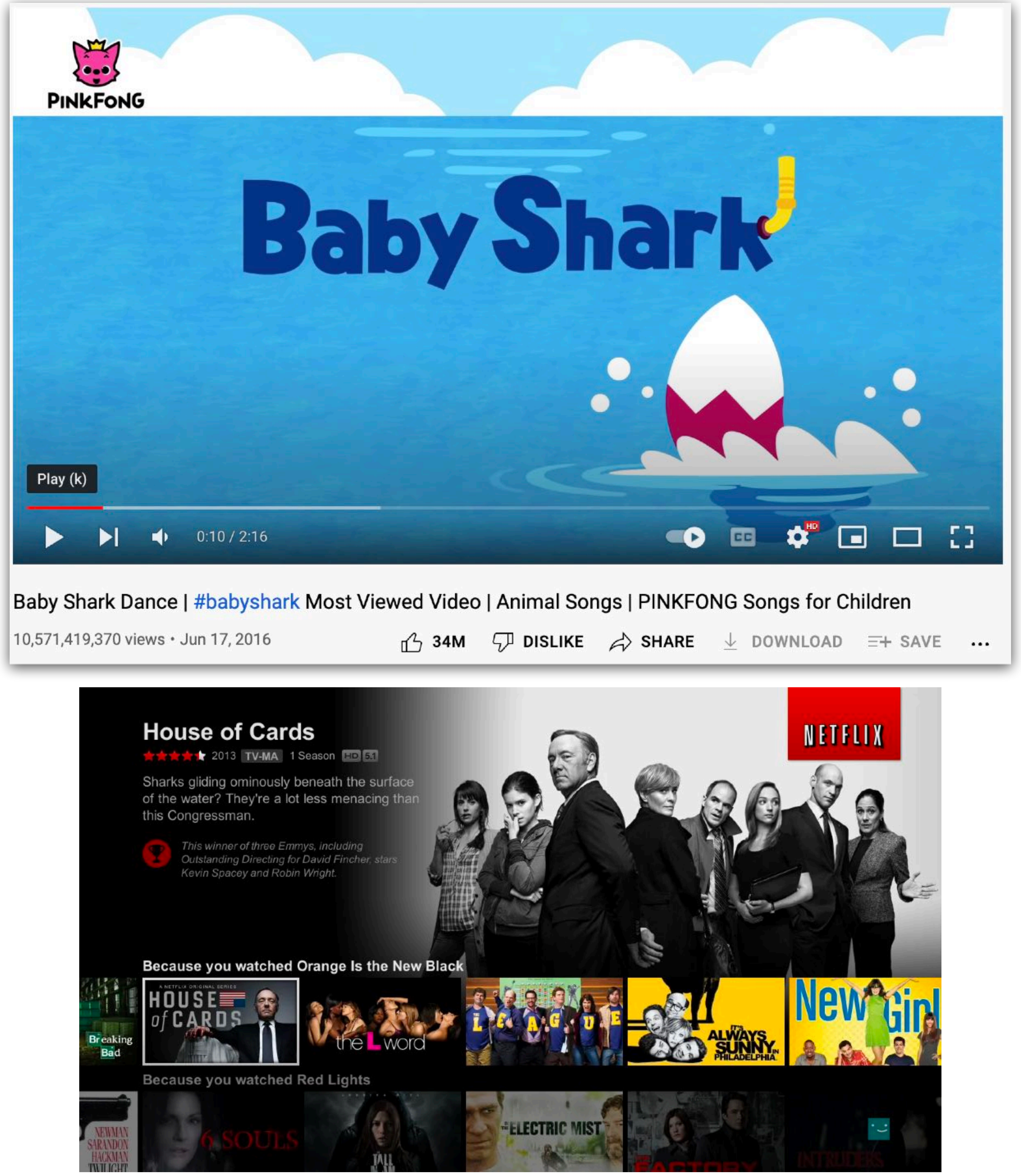
Stanford CS348K, Spring 2022

Discussion

- Google VCU paper (Ranganathan et al. 2021)

Types of video we watch on the internet

Very different latency/bandwidth requirements…



Baby Shark Dance | #babyshark Most Viewed Video | Animal Songs | PINKFONG Songs for Children

10,571,419,370 views • Jun 17, 2016

House of Cards

2013 TV-MA 1 Season

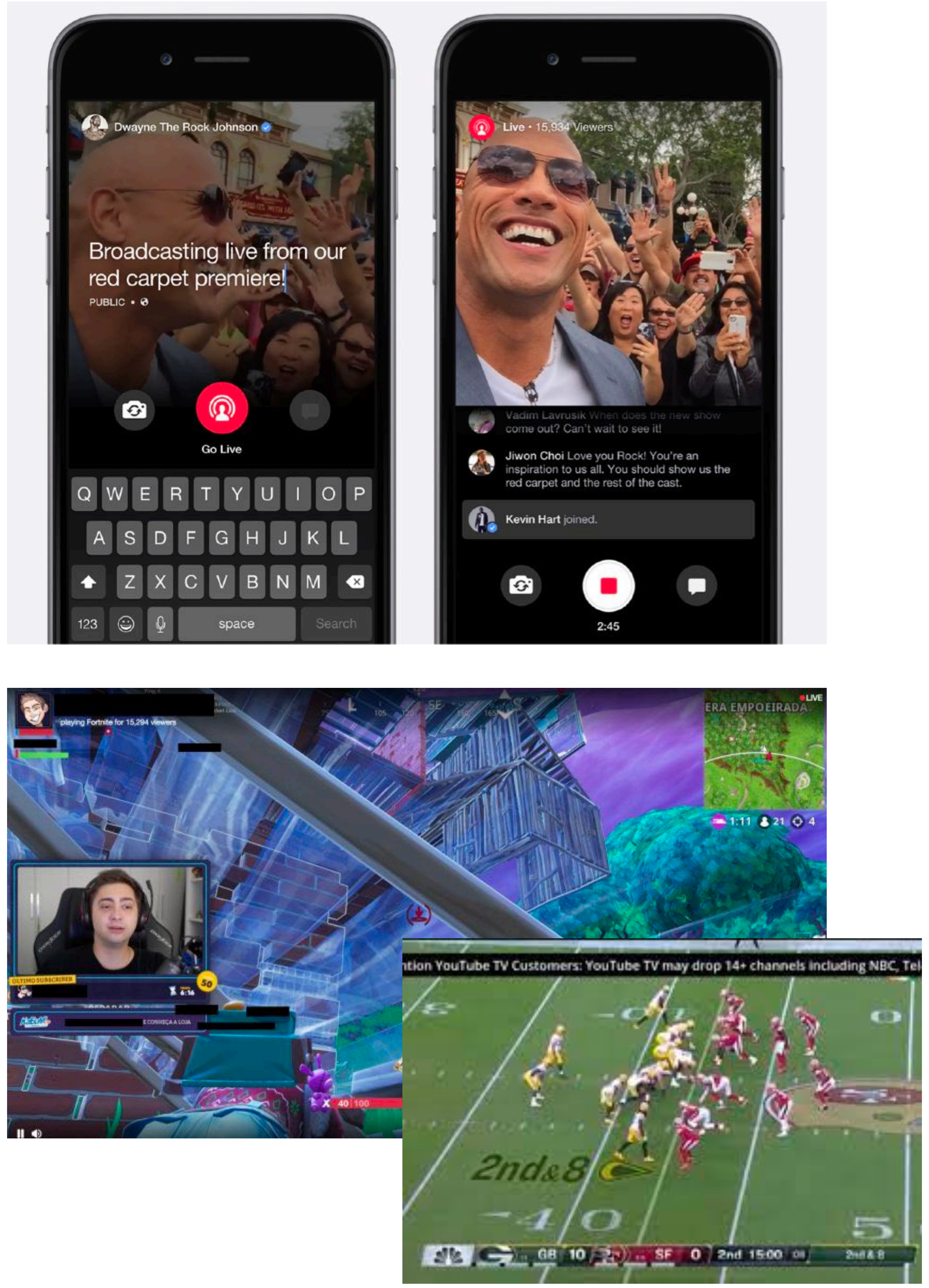
Sharks gliding ominously beneath the surface of the water? They're a lot less menacing than this Congressman.

This winner of three Emmys, including Outstanding Directing for David Fincher, stars Kevin Spacey and Robin Wright.

Because you watched Orange Is the New Black

Because you watched Red Lights

Watching videos



Broadcasting live from our red carpet premiere!

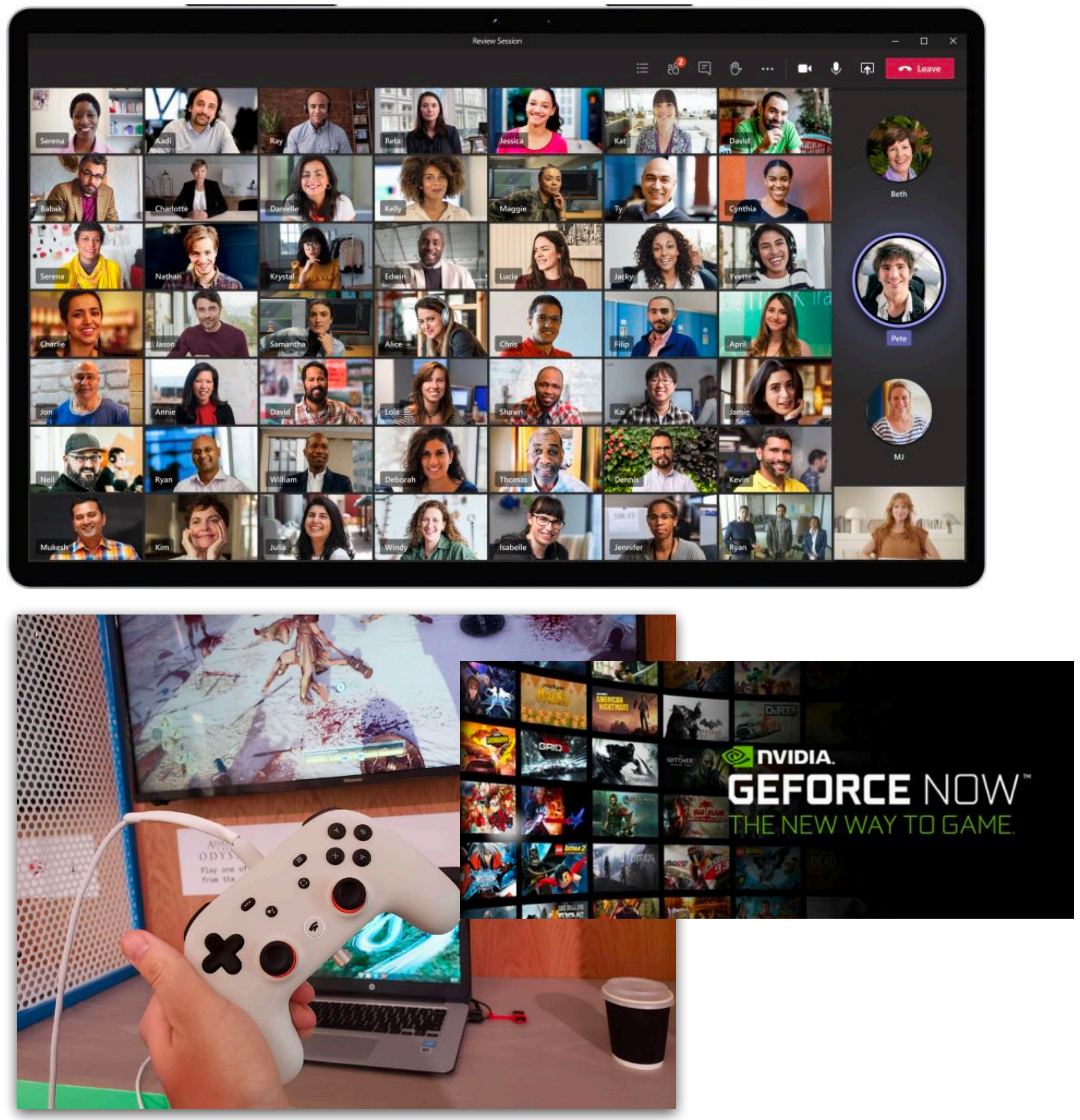
Vadim Lavrusik When does the new show come out? Can't wait to see it!

Jiwon Choi Love you Rock! You're an inspiration to us all. You should show us the red carpet and the rest of the cast.

Kevin Hart joined.

tion YouTube TV Customers: YouTube TV may drop 14+ channels including NBC, Tel

Live streaming
(Live TV, twitch, personal live streams)

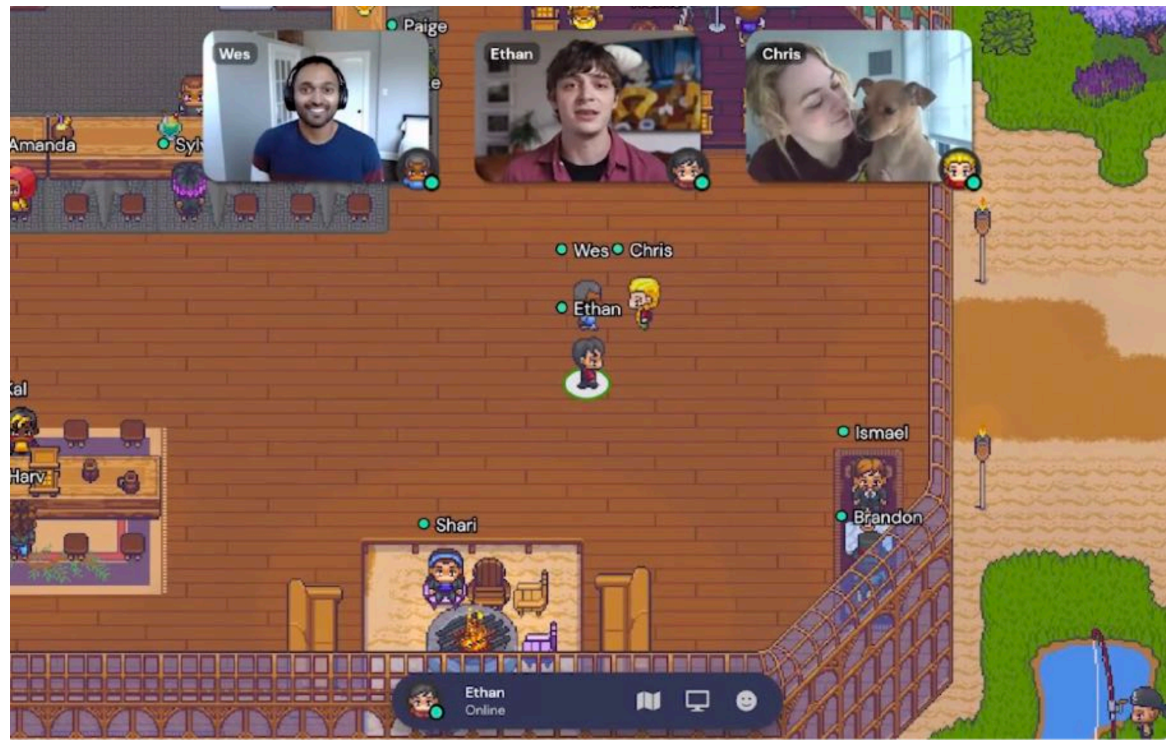


Videoconferencing,
Streaming gaming

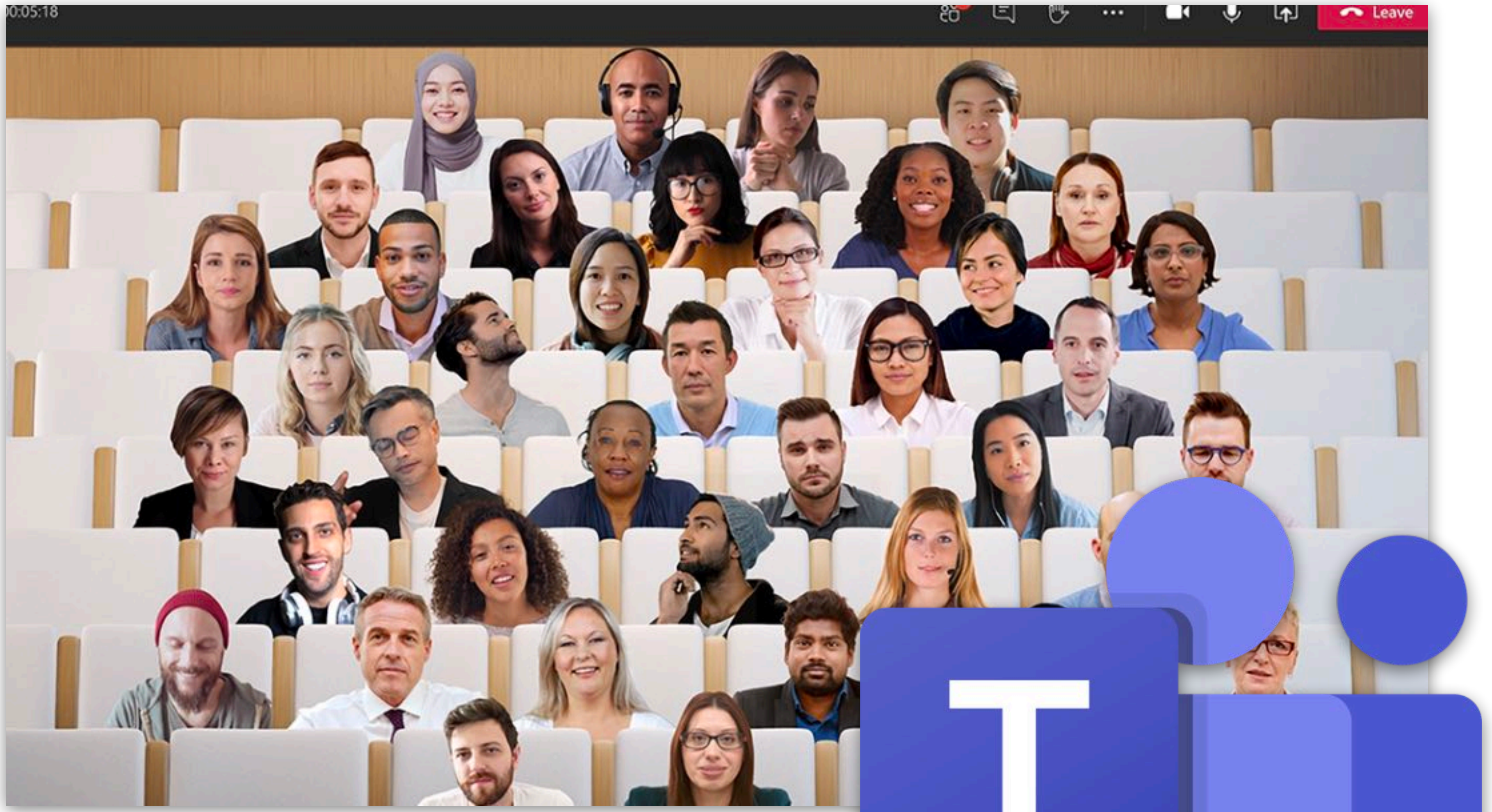
Videoconferencing systems

As you can imagine, a lot of modern interest in video conferencing(big and small!)

zoom

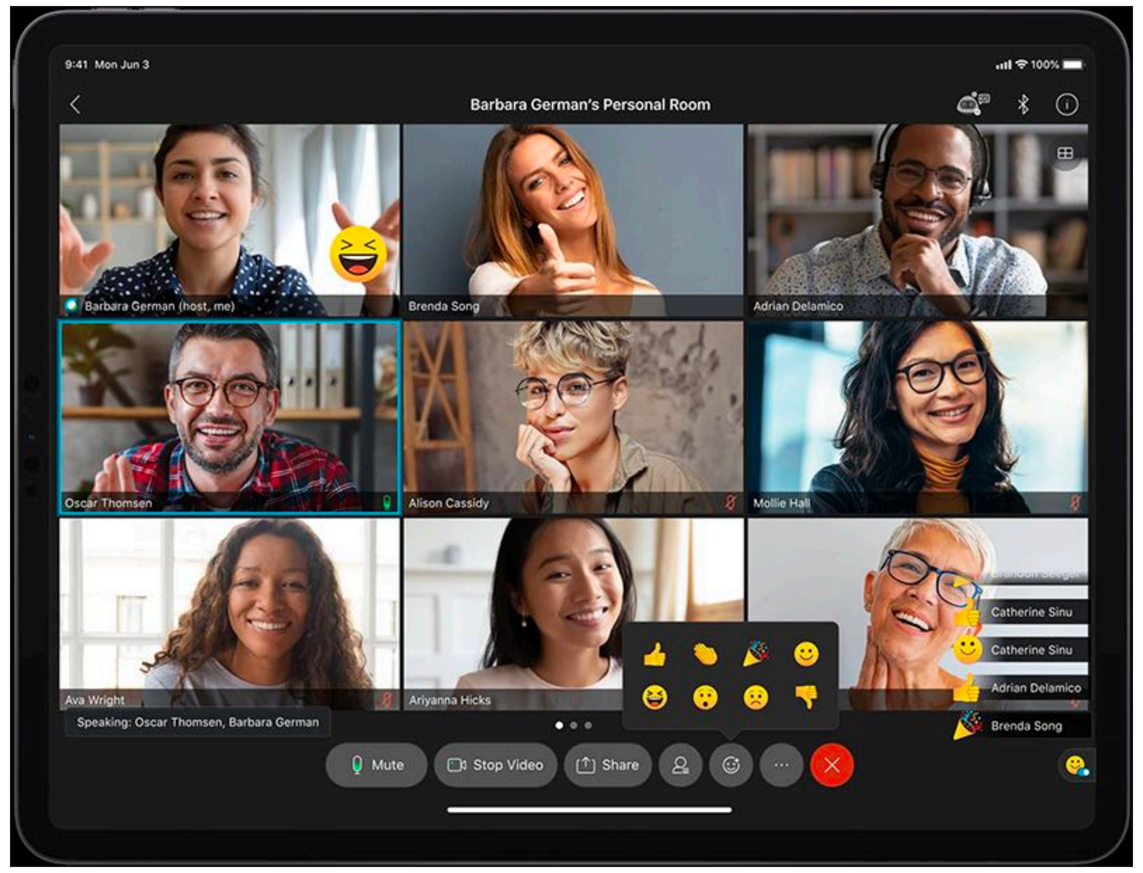


Gather



BlueJeans

mmhmm



nooks

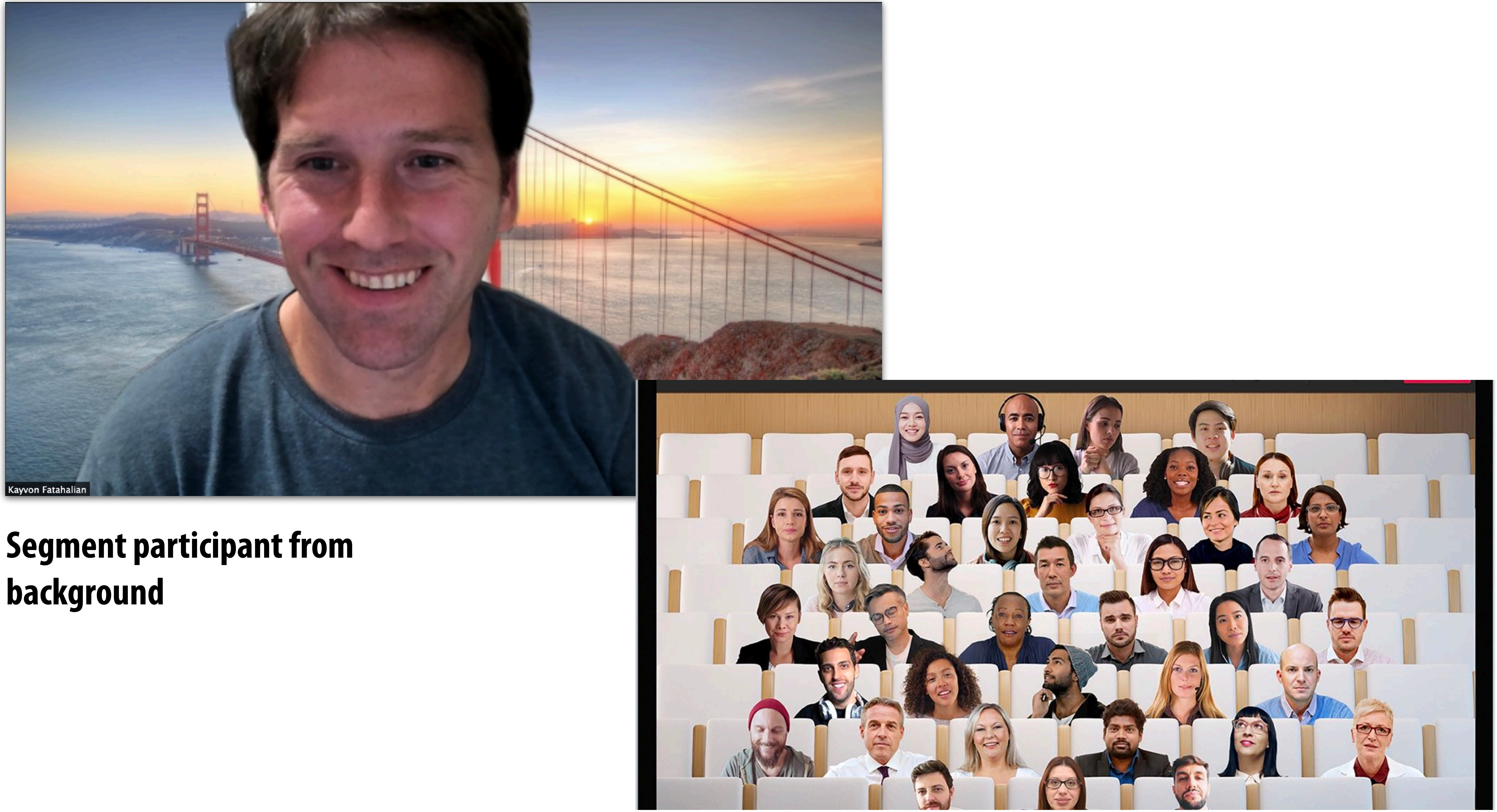
ohyay

Webex Meetings

Let’s design a video conferencing system

We want to deliver a visually rich experience similar to features of modern platforms

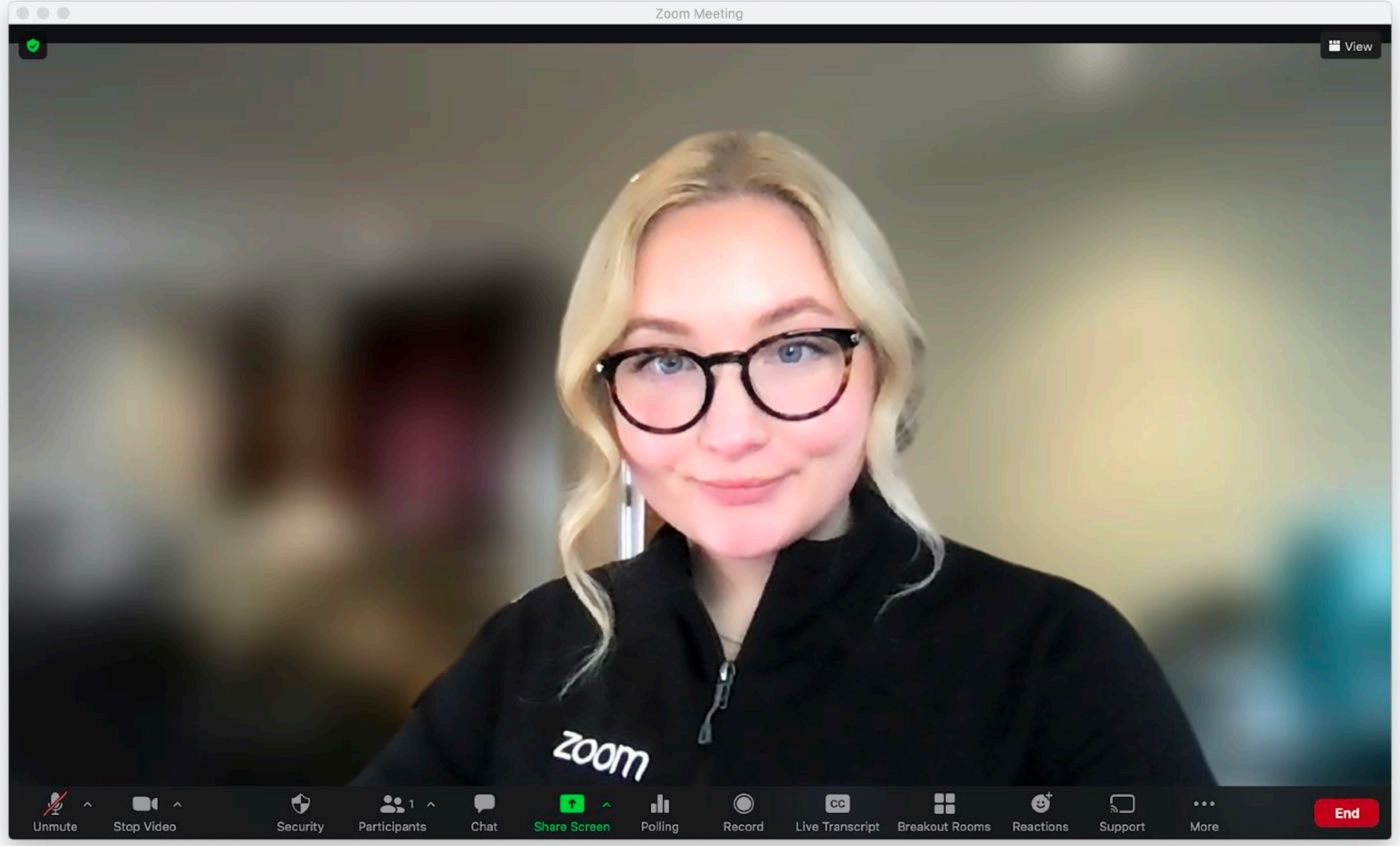
Let’s design a video conferencing system



Segment participant from background

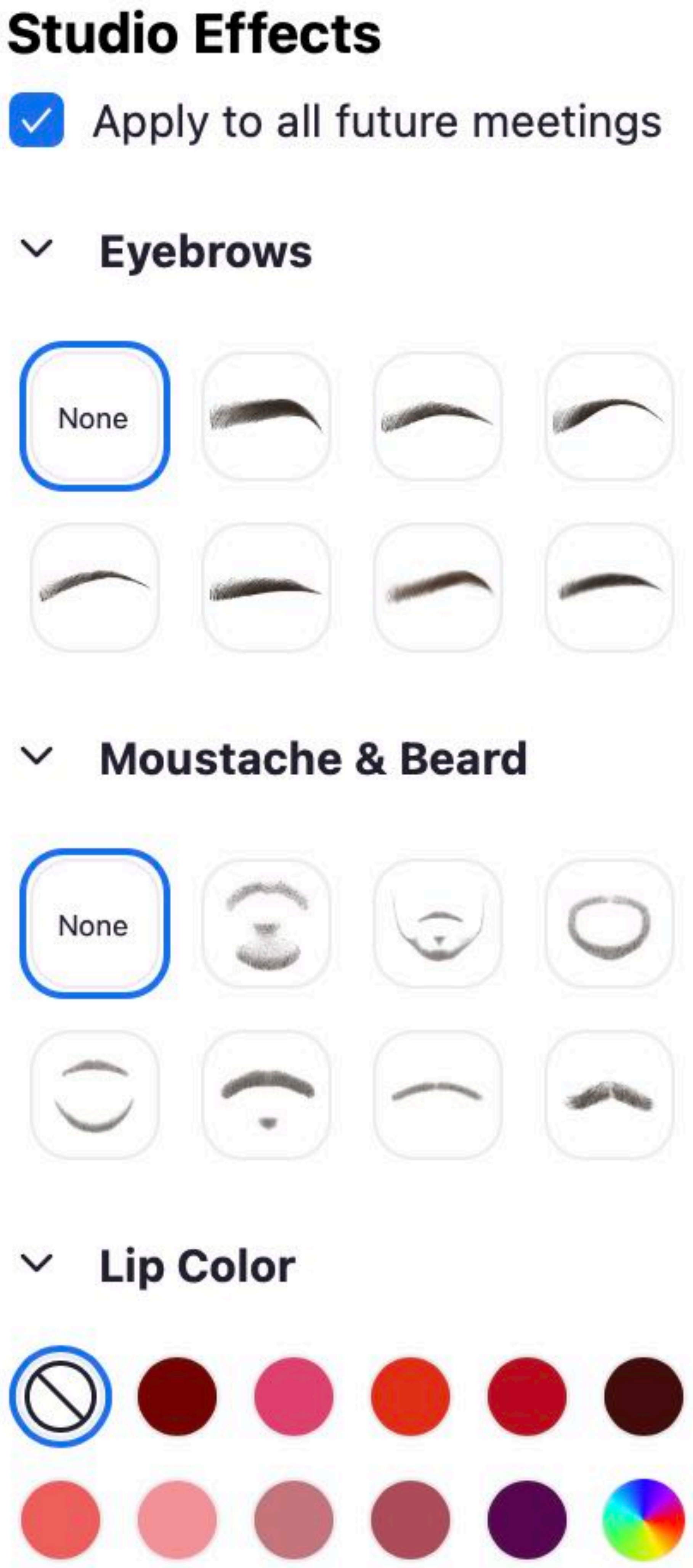
Let’s design a video conferencing system

Perform image processing to enhance look of video feed



Blur background

Render additional content



Studio Effects

☑ Apply to all future meetings

Eyebrows

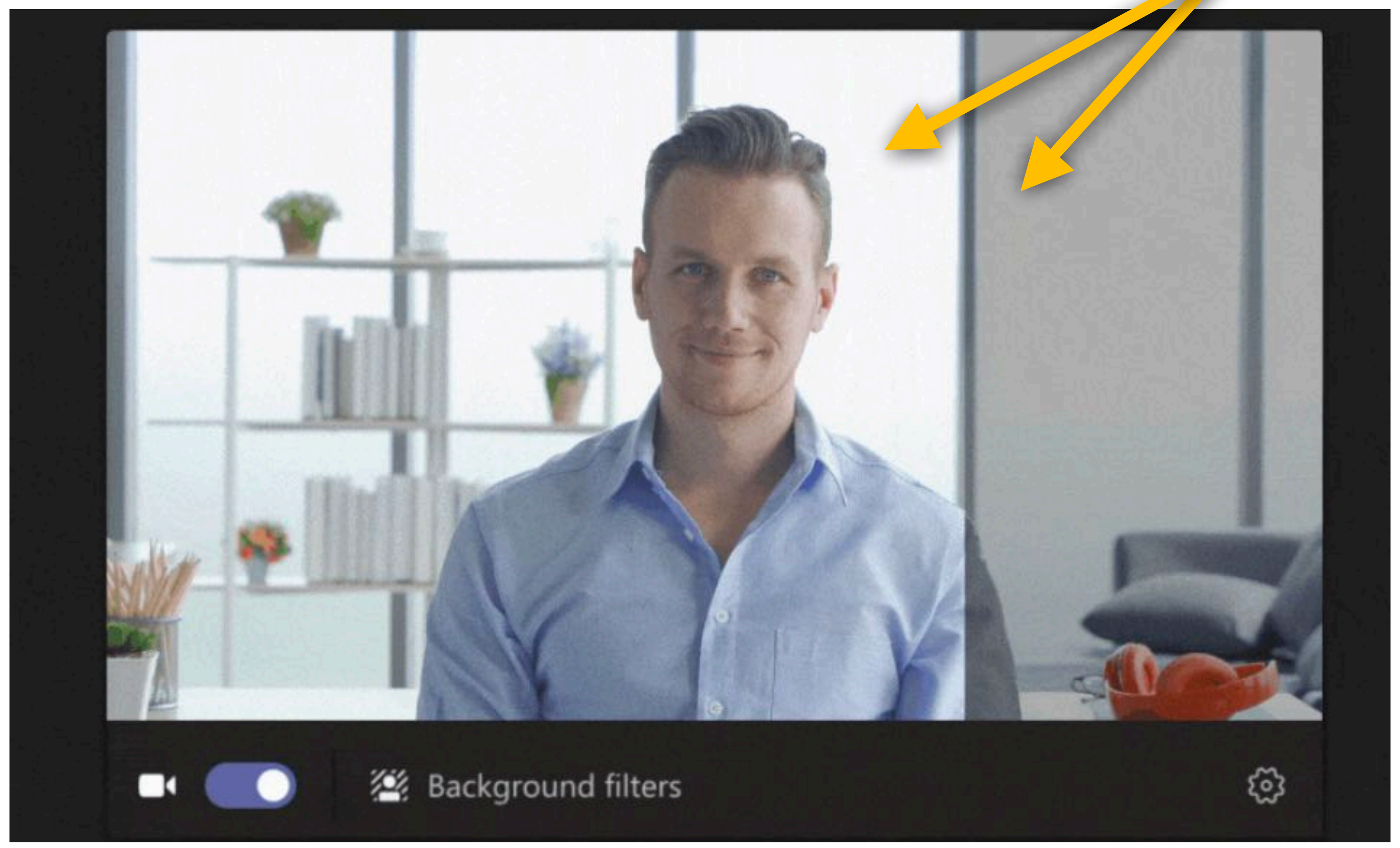
None

Moustache & Beard

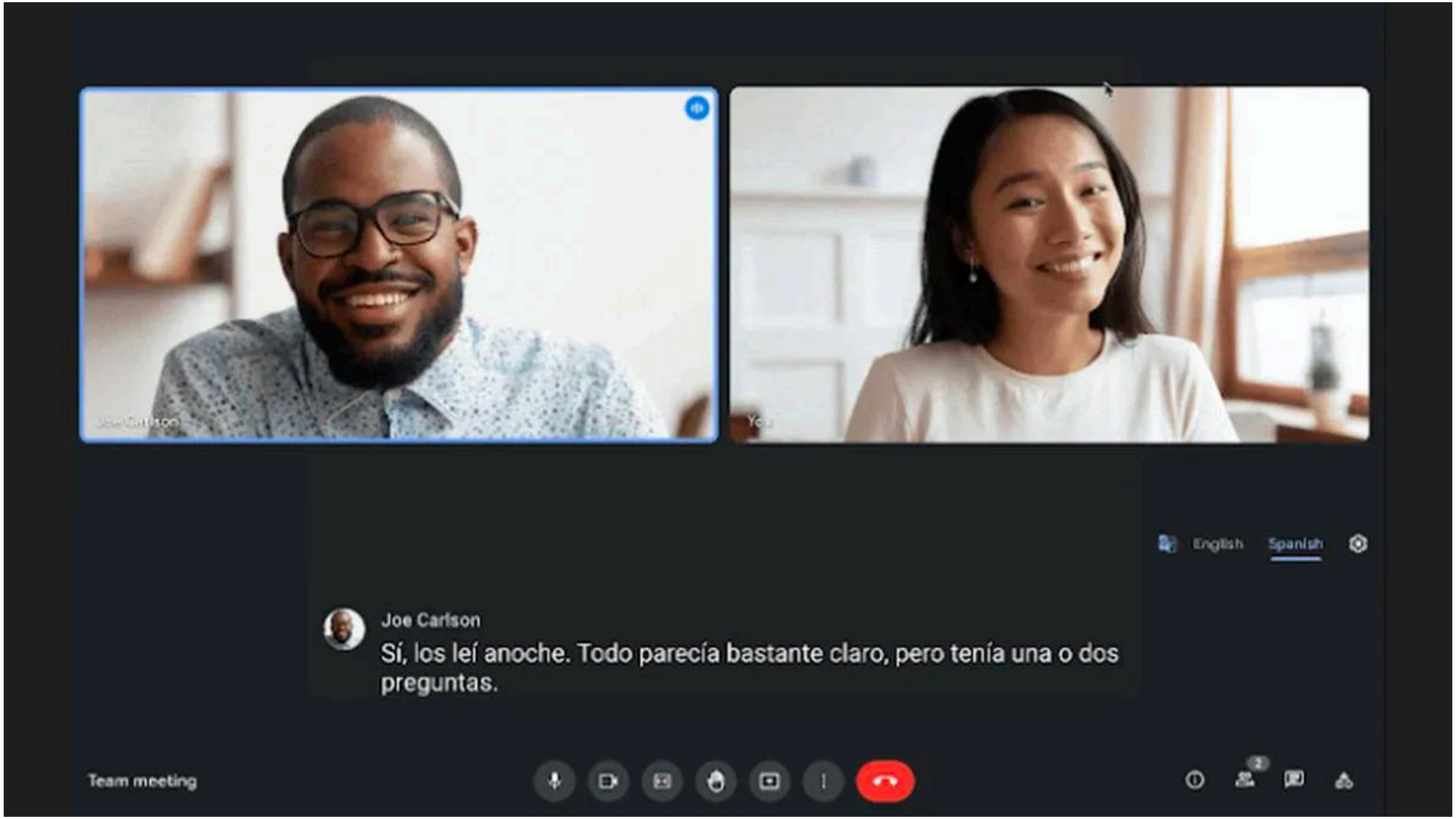
None

Lip Color

Adjust lighting



Other forms of augmentation



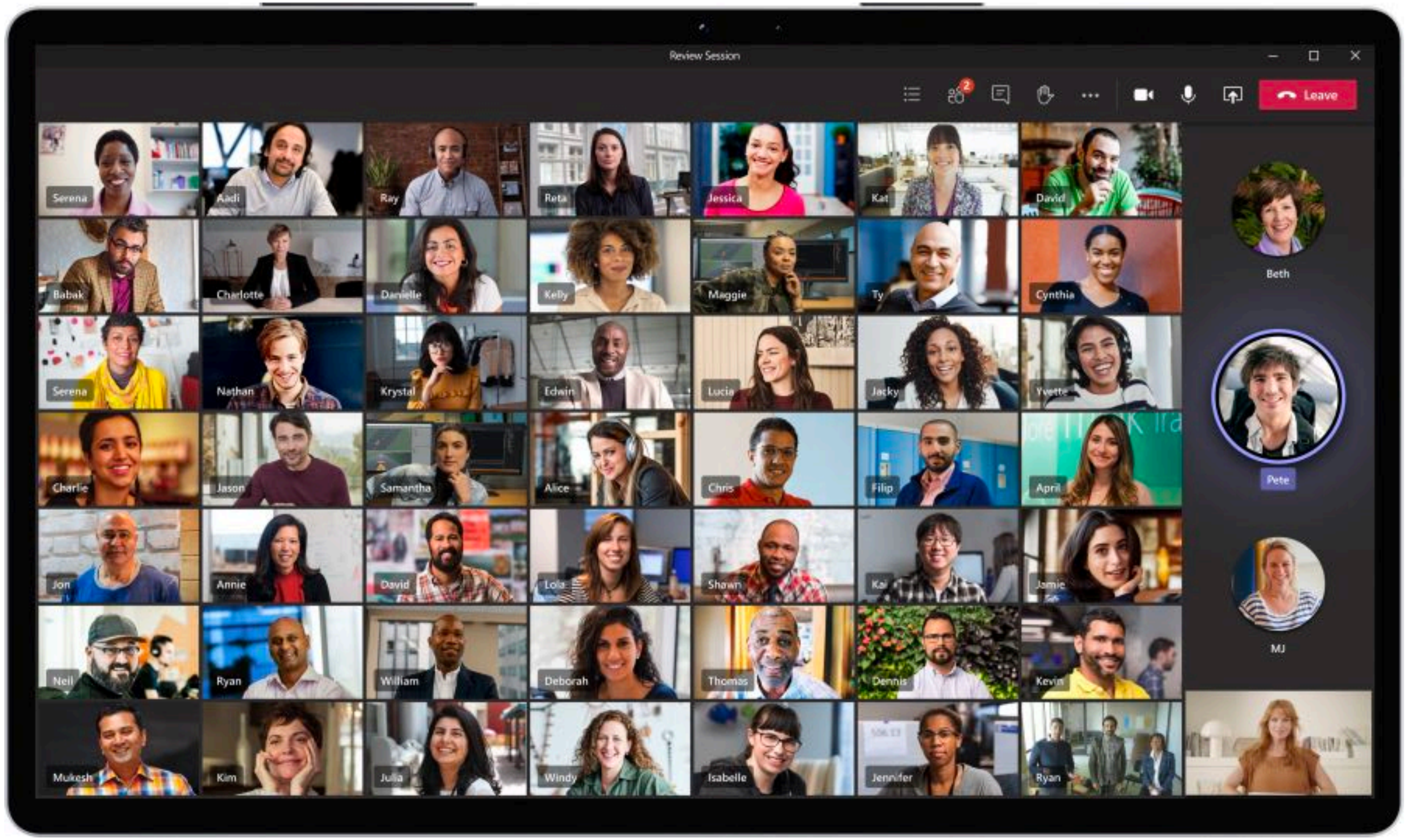
Joe Carlson

Sí, los leí anoche. Todo parecía bastante claro, pero tenía una o dos preguntas.

Real-time translation and captioning

Let’s design a video conferencing system

Large gallery views: companies raced to provide 7x7 gallery in 2020



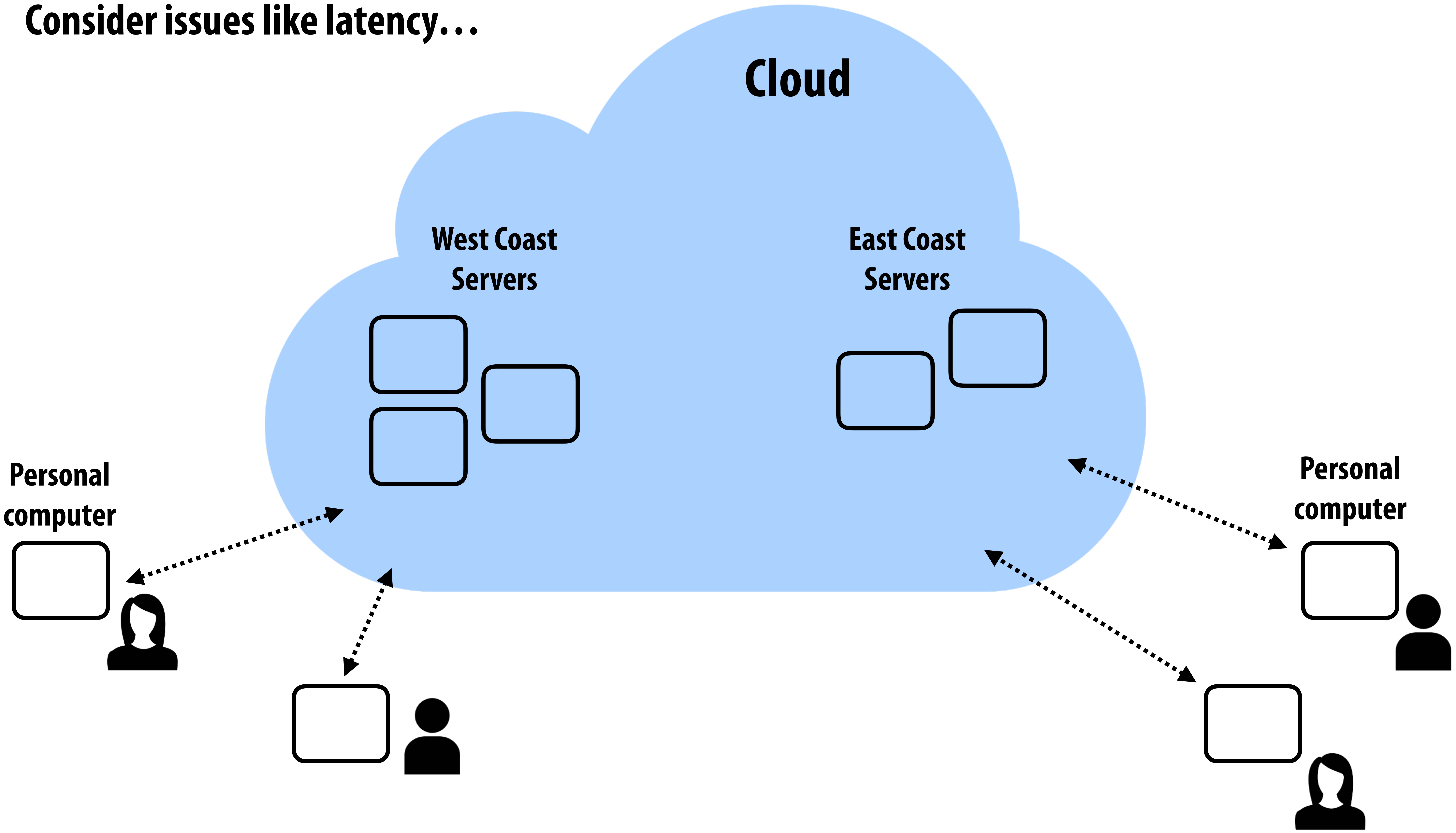
Maximum participants displayed per screen in Gallery View:

○ 25 participants ● 49 participants

Deliver to wide range of clients and network settings

Setup…

Consider issues like latency…

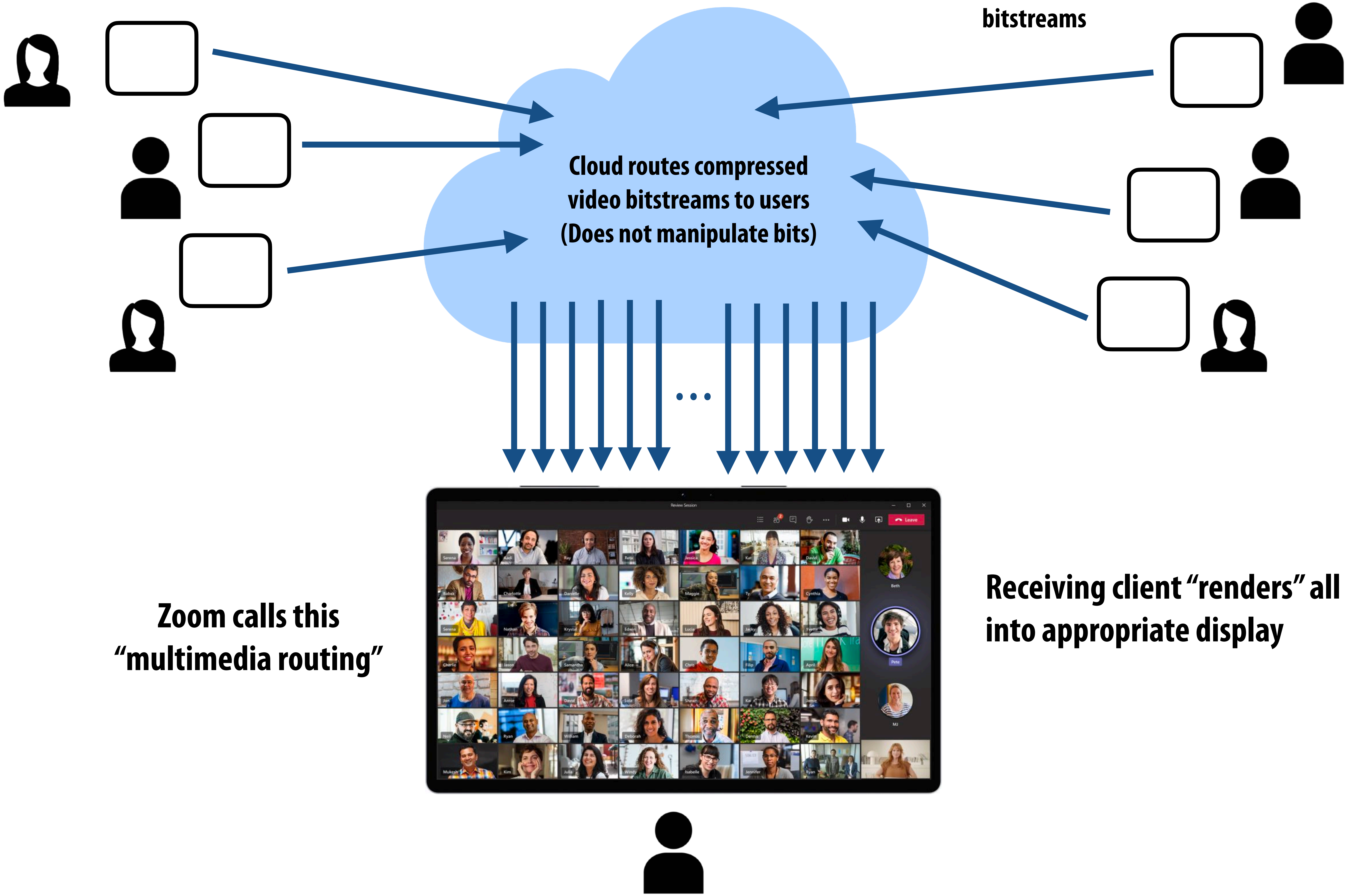


Icon credits: person by mim studio from the Noun Project, avatar by Soremba from the Noun Project

Q. Should we transcode/process video on our cloud servers?

- What are advantages (to users? To us the provider)?
- What are disadvantages?

Implementing gallery view



Clients transmit individually compressed bitstreams

Cloud routes compressed video bitstreams to users (Does not manipulate bits)

Zoom calls this “multimedia routing”

Receiving client “renders” all streams into appropriate display

One drawback of this design

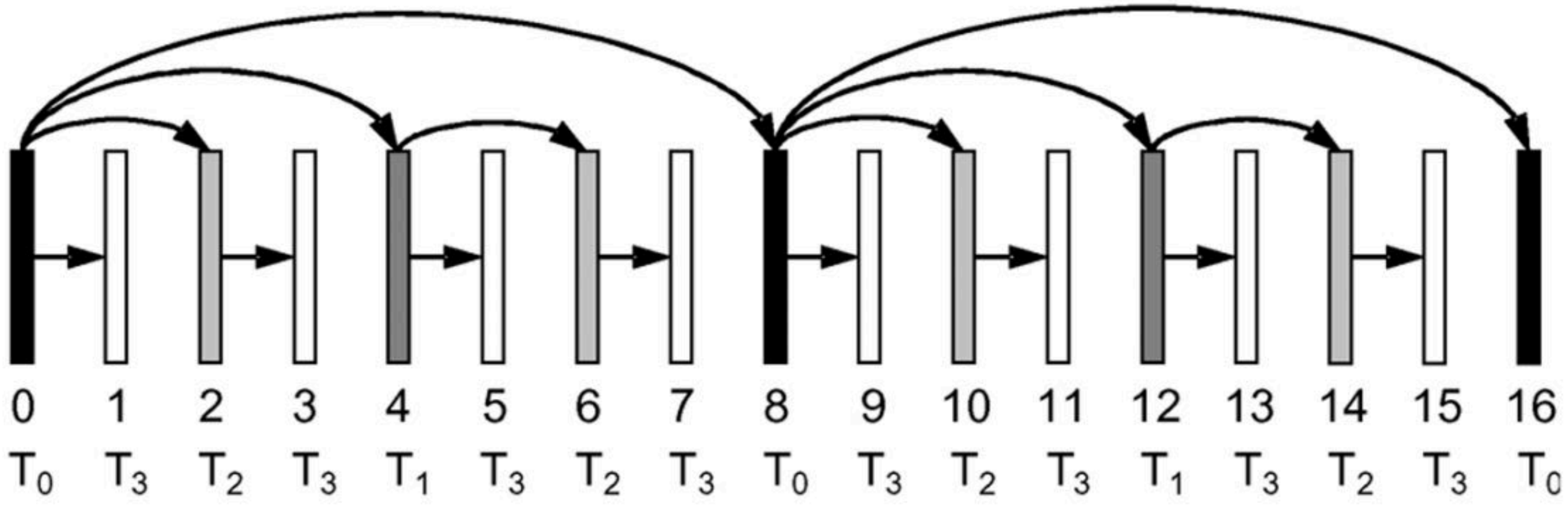
- If each client is providing a single compressed video stream, that means each person on the video call must receive the same bits right? (What if they are on different network connections?)

Scalable video codec (SVC)

SVC is an extension of H.264 standard

- "Scalable" compressed video bitstream: subsets of the bitstream encode valid video streams for a decoder
  - Implication: if packets get lost, the remaining packets form a valid H.264 bitstream, albeit at lower resolution or quality

Example: temporal scalability


0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
T_0 T_3 T_2 T_3 T_1 T_3 T_2 T_3 T_0 T_3 T_2 T_3 T_1 T_3 T_2 T_3 T_0

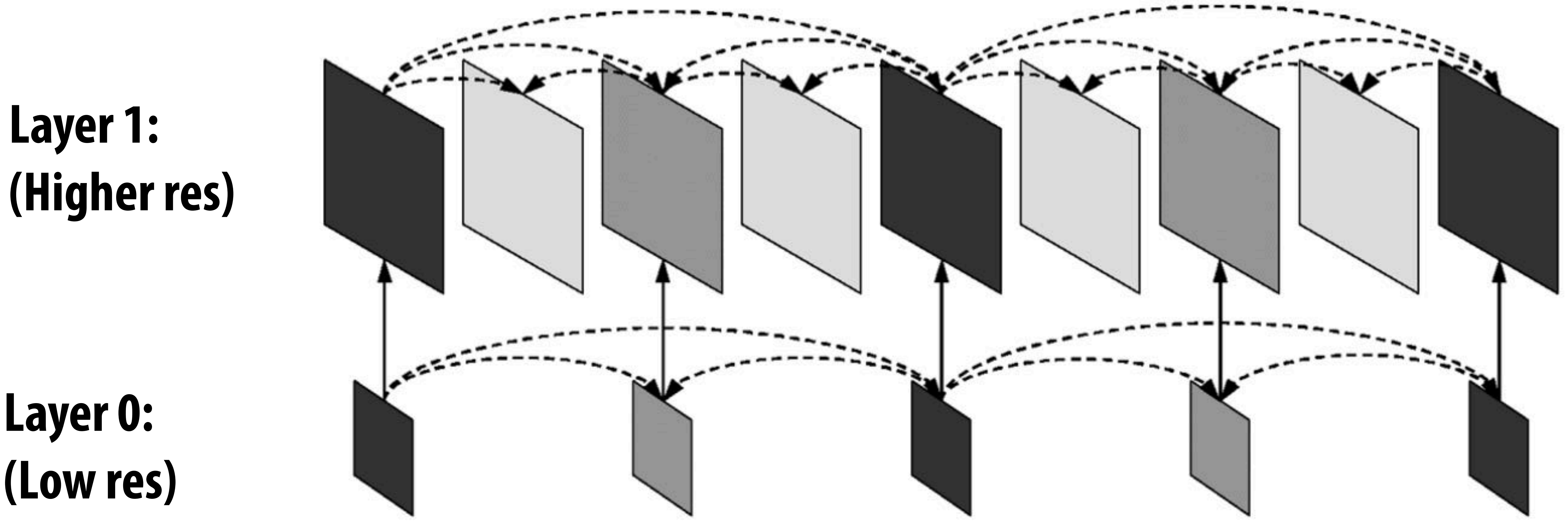
Layer 0: (T_0) defines valid video at frame rate R
Layer 1 (T_1) defines bumps frame rate to 2R
...

Note how layer 0 information is used to predict higher layer information

Scalable video codec (SVC)

SVC is an extension of H.264 standard

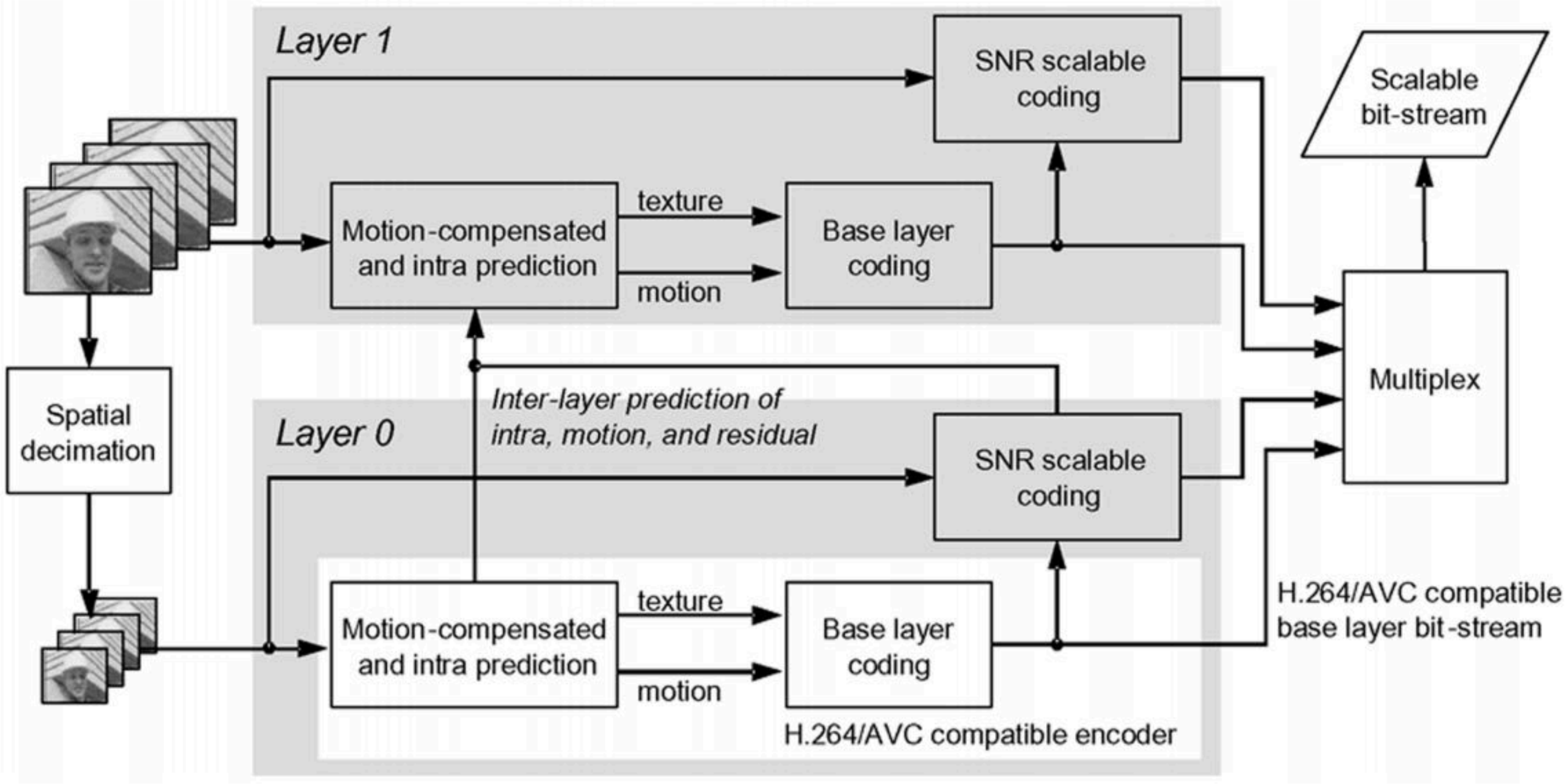
Example: spatial scalability



Again, note how layer 0 information is used to predict higher layer information
(Higher efficiency than independently encoding two video streams)

Layer 0: defines valid video at low resolution (and low frame rate)
Layer 1: provides additional information for higher resolution (and higher frame rate) video

Scalable video codec (SVC) encoder

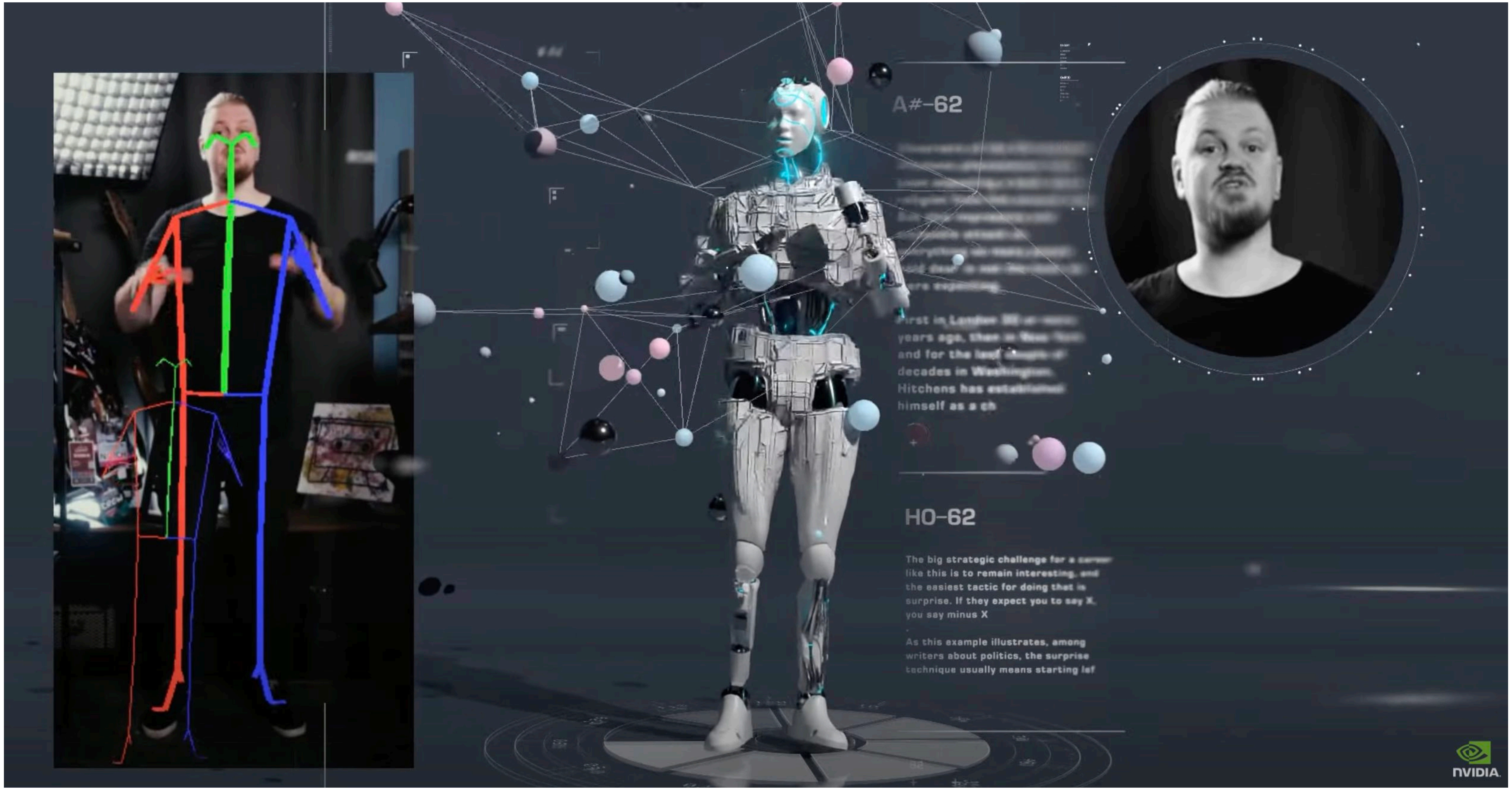

Layer 1
SNR scalable coding
Scalable bit-stream
Motion-compensated and intra prediction
texture
motion
Base layer coding
Multiplex
Spatial decimation
Inter-layer prediction of intra, motion, and residual
Layer 0
SNR scalable coding
Motion-compensated and intra prediction
texture
motion
Base layer coding
H.264/AVC compatible base layer bit-stream
H.264/AVC compatible encoder

Costs: higher encoding/decoding costs

(But possible on modern clients as SVC is supported in hardware)

NVIDIA Maxine

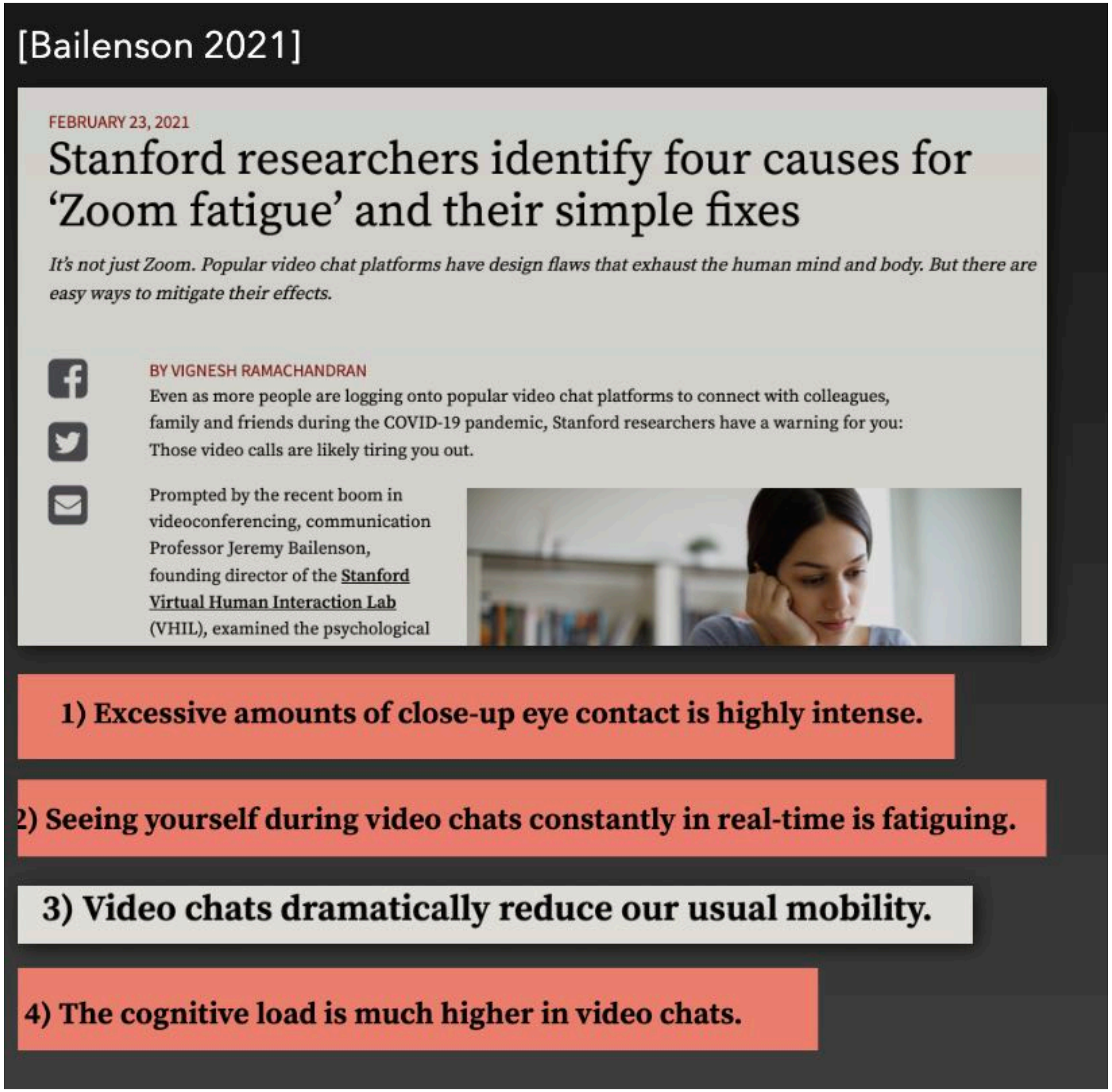
GPU-accelerated video processing for video conferencing applications



Examples: avatar control, video superresolution, advanced background segmentation

What do people *really need*?

“Zoom fatigue” is very real



[Bailenson 2021]

FEBRUARY 23, 2021

Stanford researchers identify four causes for ‘Zoom fatigue’ and their simple fixes

It’s not just Zoom. Popular video chat platforms have design flaws that exhaust the human mind and body. But there are easy ways to mitigate their effects.

BY VIGNESH RAMACHANDRAN

Even as more people are logging onto popular video chat platforms to connect with colleagues, family and friends during the COVID-19 pandemic, Stanford researchers have a warning for you: Those video calls are likely tiring you out.

Prompted by the recent boom in videoconferencing, communication Professor Jeremy Bailenson, founding director of the Stanford Virtual Human Interaction Lab (VHIL), examined the psychological

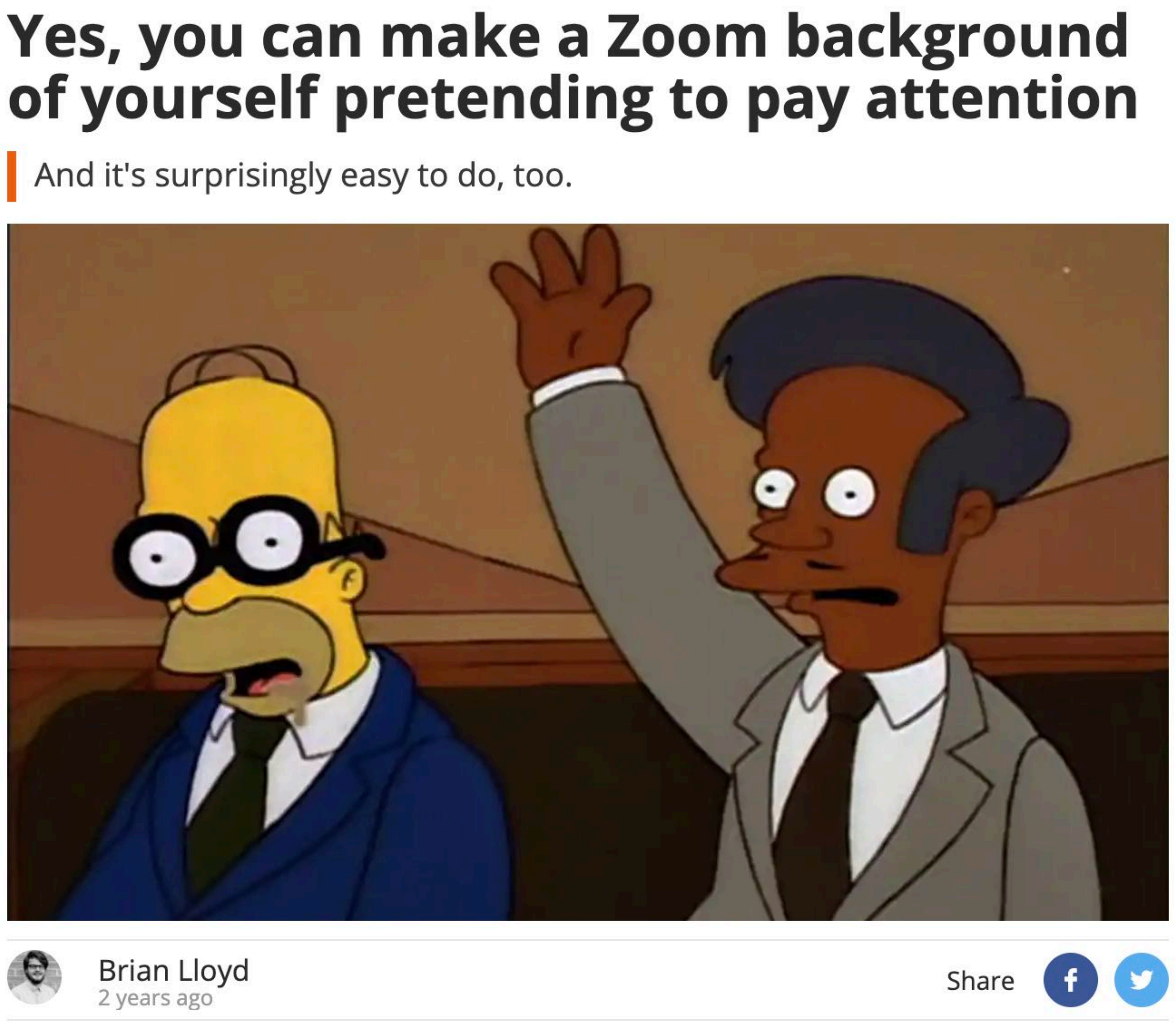
1) Excessive amounts of close-up eye contact is highly intense.

2) Seeing yourself during video chats constantly in real-time is fatiguing.

3) Video chats dramatically reduce our usual mobility.

4) The cognitive load is much higher in video chats.

The best camera is the one that's off?

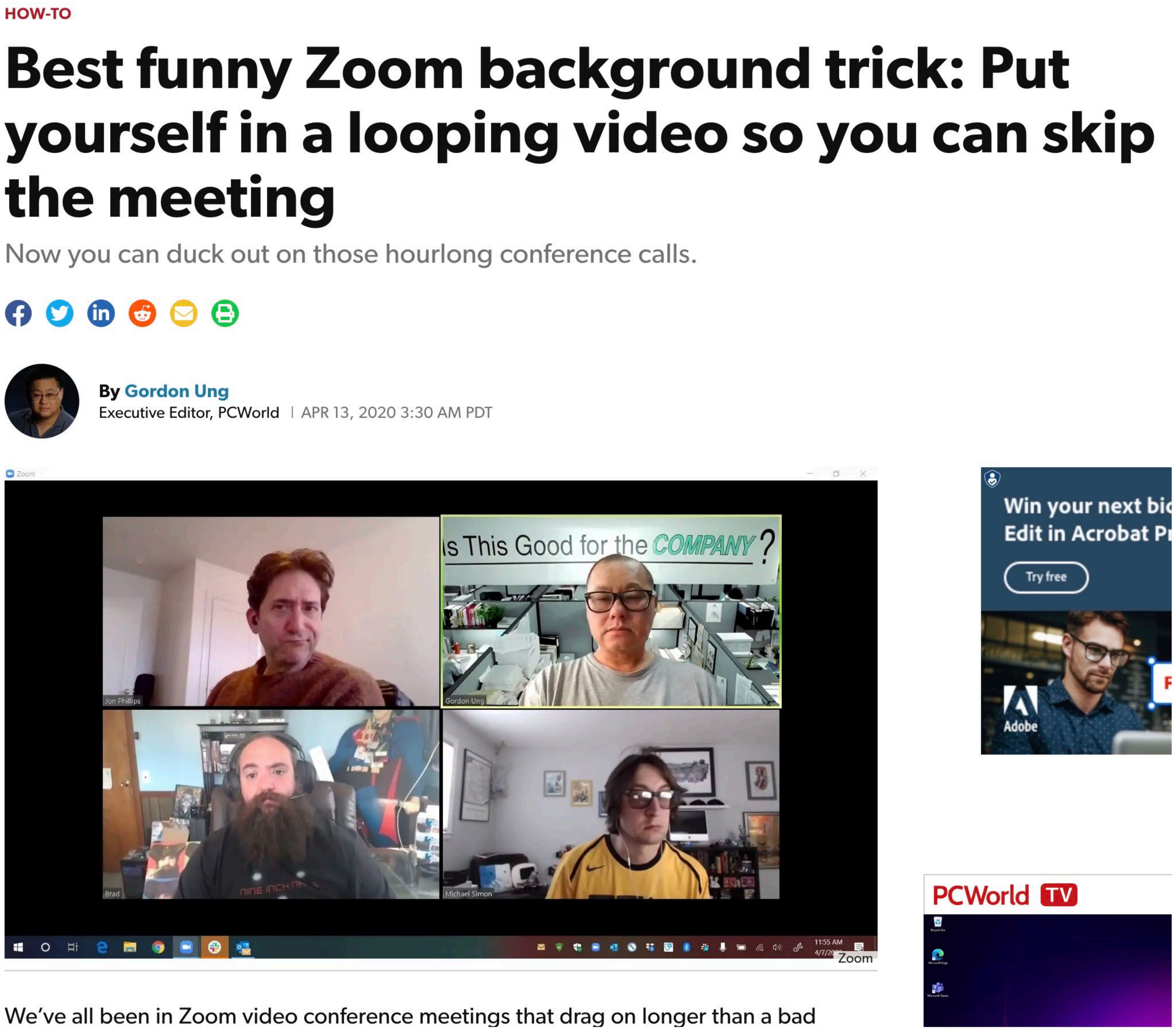


Yes, you can make a Zoom background of yourself pretending to pay attention

And it's surprisingly easy to do, too.

Brian Lloyd
2 years ago

Share



HOW-TO

Best funny Zoom background trick: Put yourself in a looping video so you can skip the meeting

Now you can duck out on those hourlong conference calls.

By Gordon Ung
Executive Editor, PCWorld | APR 13, 2020 3:30 AM PDT

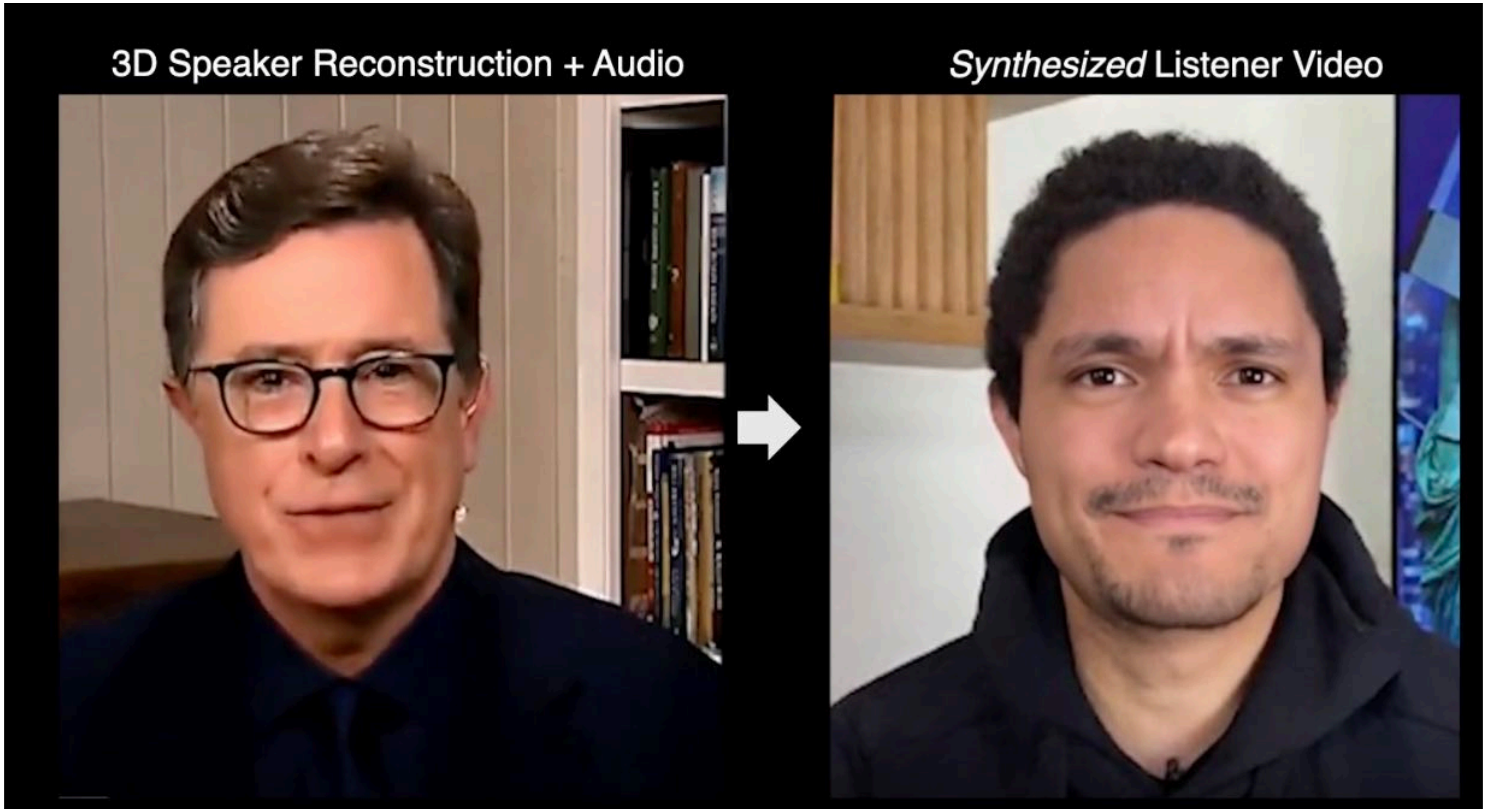
We've all been in Zoom video conference meetings that drag on longer than a bad

PCWorld TV

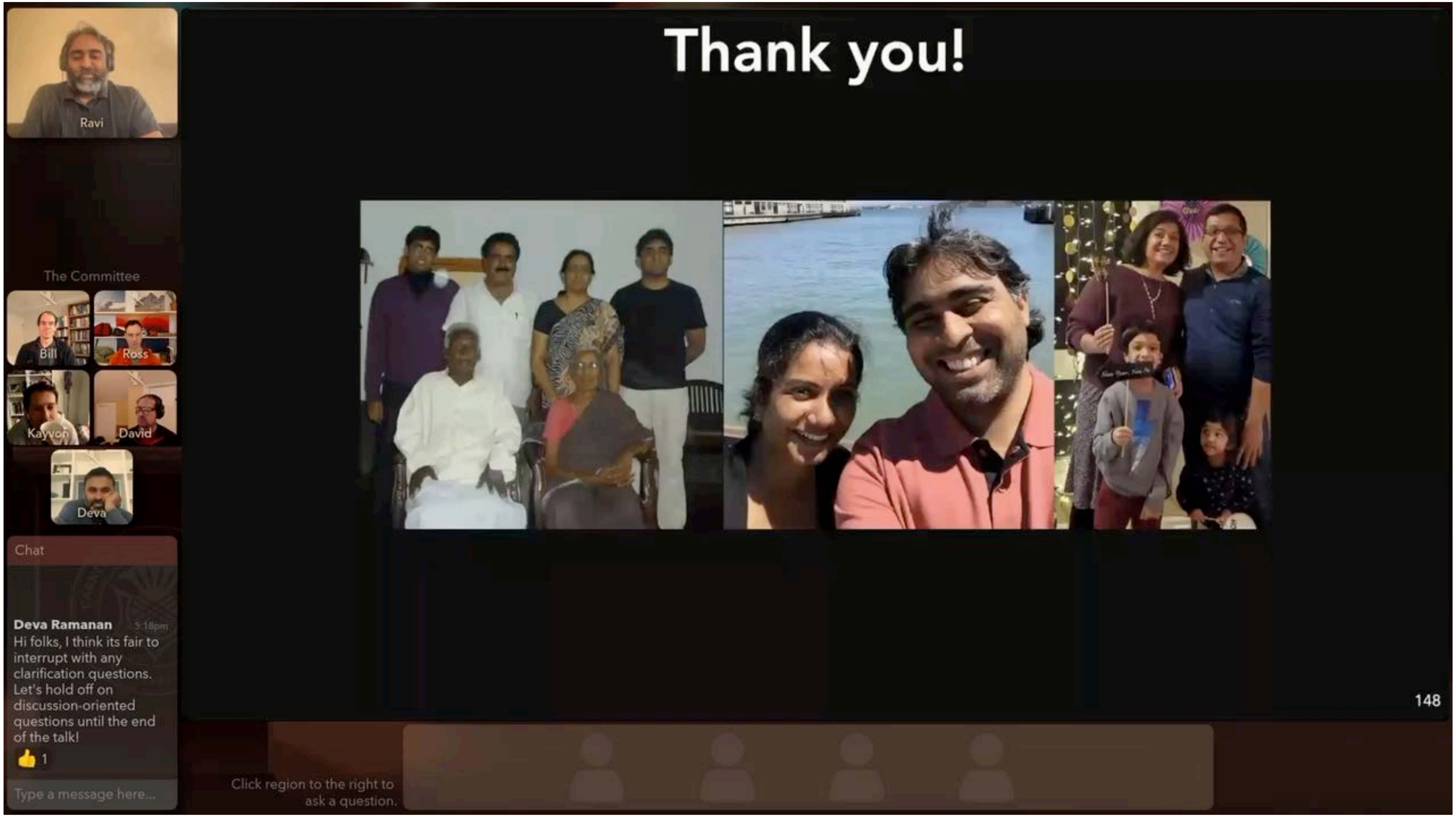
Synthesizing reactions?

Input: audio of speaker

Output: video of listener's reaction


3D Speaker Reconstruction + Audio
Synthesized Listener Video

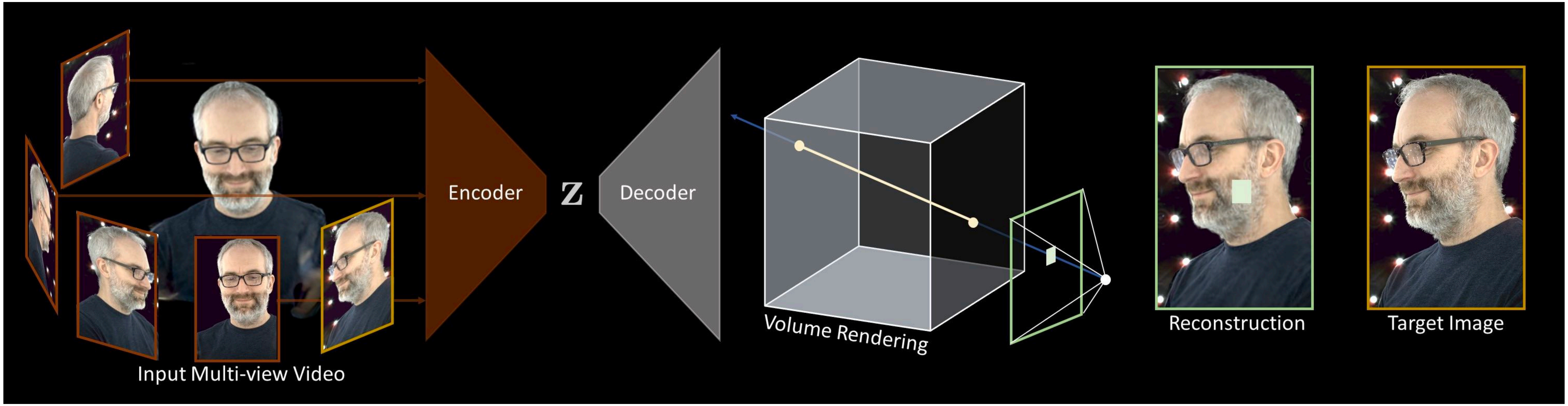
User-triggered effects (examples: audio clips, “reactions”)



Deva Ramanan 5:18pm
Hi folks, I think its fair to interrupt with any clarification questions. Let's hold off on discussion-oriented questions until the end of the talk!
👍 1

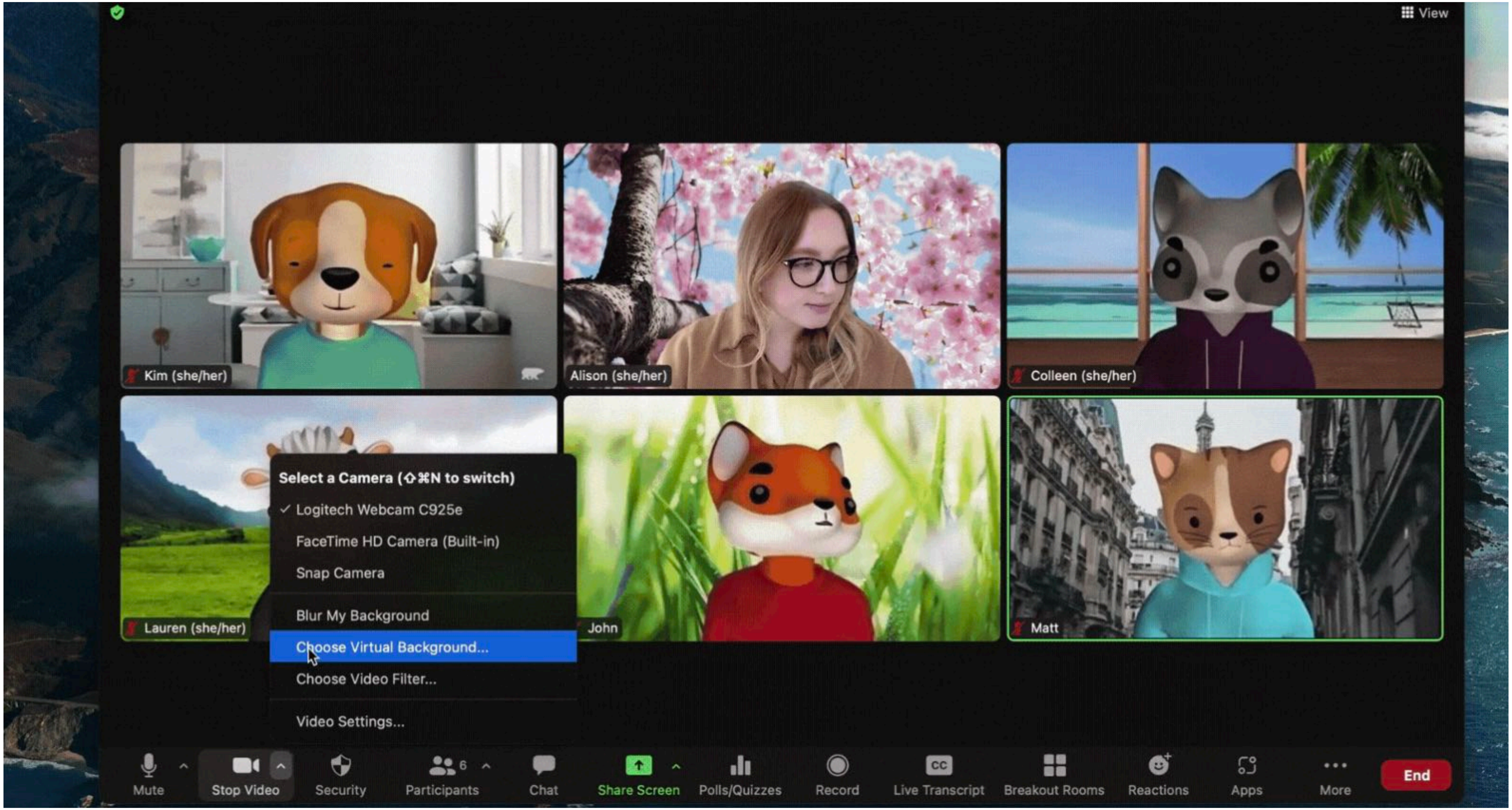
Neural volumes

- Learn to encode multiple views of a person into a latency code (z) that is decoded into a volume than can be rendered with conventional graphics techniques *from any viewpoint*

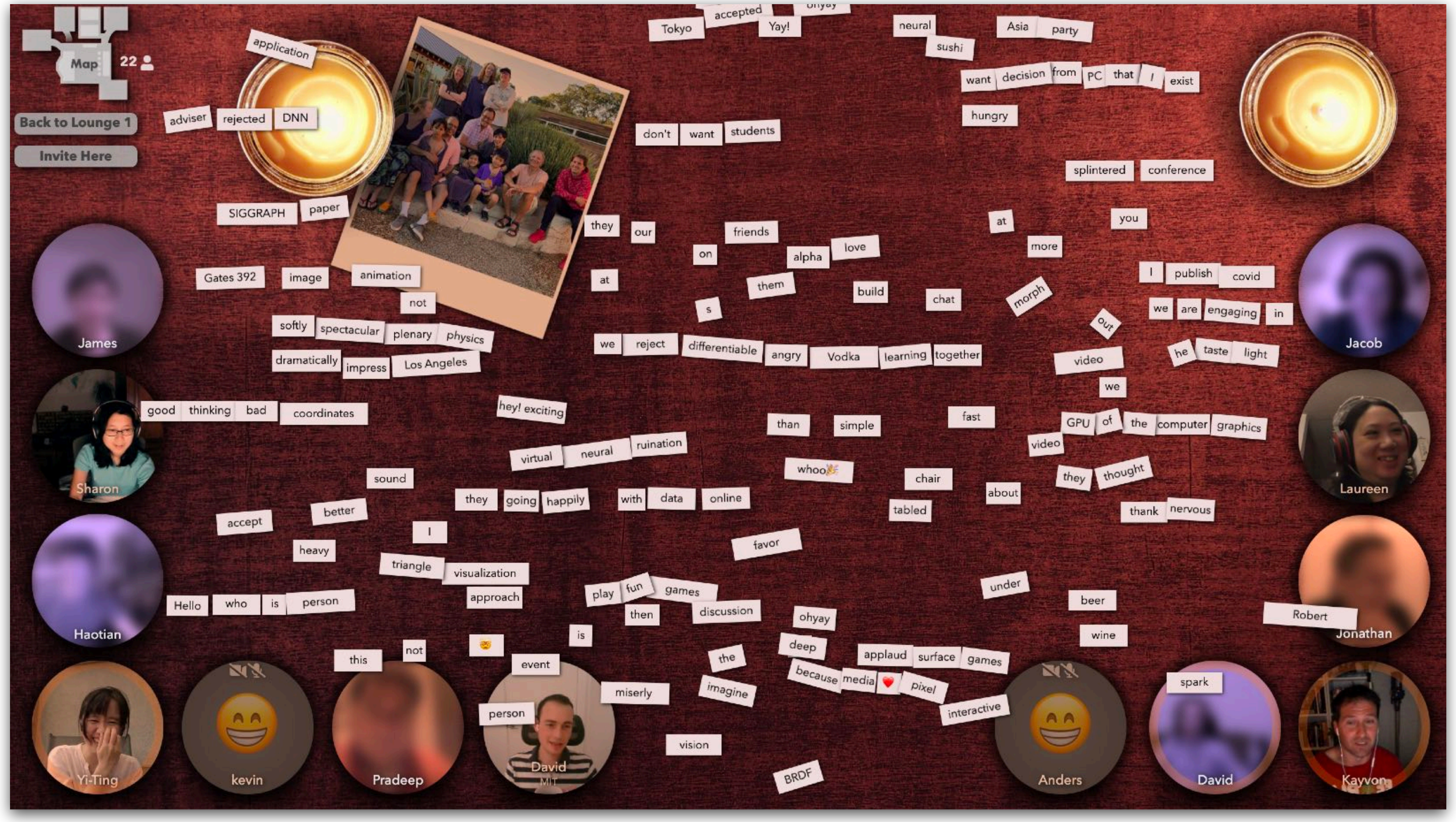


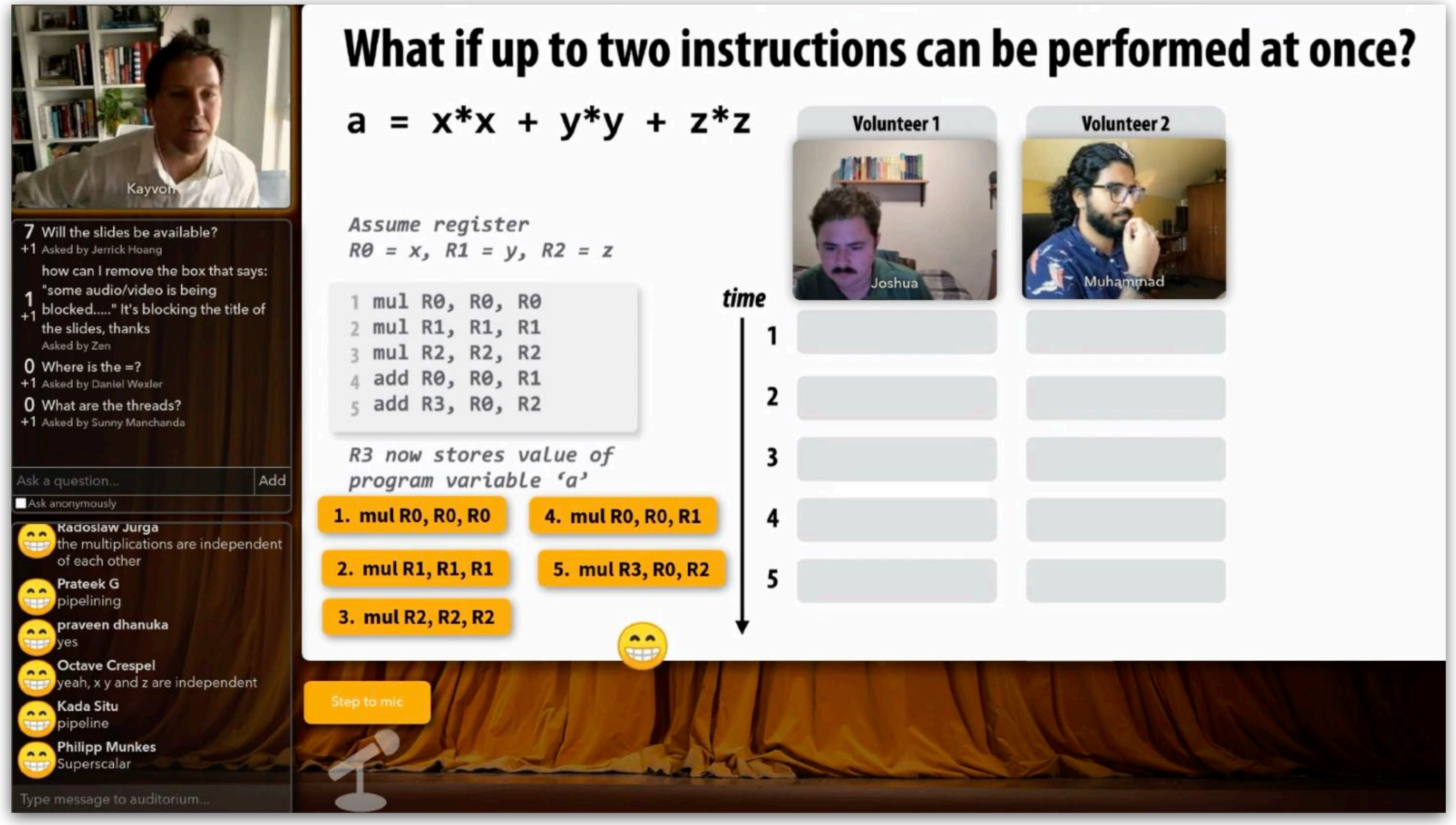
- Motivated by VR applications

Zoom avatars / Snapcam Lenses



More examples (demo)





What if up to two instructions can be performed at once?

a = x*x + y*y + z*z

Assume register
R0 = x, R1 = y, R2 = z

```
mul R0, R0, R0
mul R1, R1, R1
mul R2, R2, R2
add R0, R0, R1
add R3, R0, R2
```

R3 now stores value of program variable ‘a’

1. mul R0, R0, R0
2. mul R1, R1, R1
3. mul R2, R2, R2
4. mul R0, R0, R1
5. mul R3, R0, R2

7 +1 Will the slides be available? Asked by Jerrick Hoang

1 +1 how can I remove the box that says: "some audio/video is being blocked....." It's blocking the title of the slides, thanks Asked by Zen

0 +1 Where is the =? Asked by Daniel Wexler

0 +1 What are the threads? Asked by Sunny Manchanda

Ask a question... Add

□ Ask anonymously

Radoslaw Jurga the multiplications are independent of each other

Prateek G pipelining

praveen dhanuka yes

Octave Crespel yeah, x y and z are independent

Kada Situ pipeline

Philipp Munkes Superscalar

Type message to auditorium...

Discussion:

Where is the ethical line between “augmenting or abstracting what’s real” and “fake”?

How can technology strike a balance between facilitating better forms of communication + a sense of presence (e.g., working from home) vs. ensuring privacy and personal space?

Can you think of ways where widespread use of near-photorealistic digital personals (e.g., for work calls) might lead to unexpected harms?

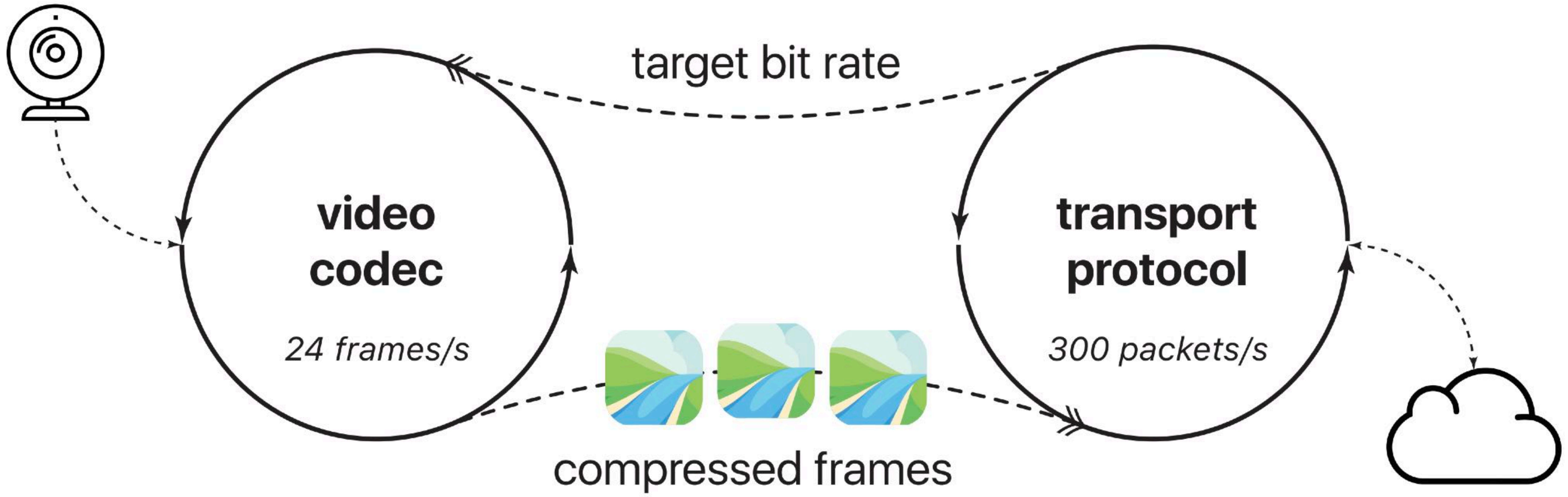
Do virtual meetings promote more diverse representation? Set it back?

(If time)
Co-designing video compressor and network transport

[Credit: Fouladi et al. 2018]

Status quo

- Video encoder proceeds to compress video frames, targeting a bit rate (on average) provided by the network protocol
- But any one frame may be too large or small (some may be hard to predict)



Encoder: targets an average bit rate (bits/second)

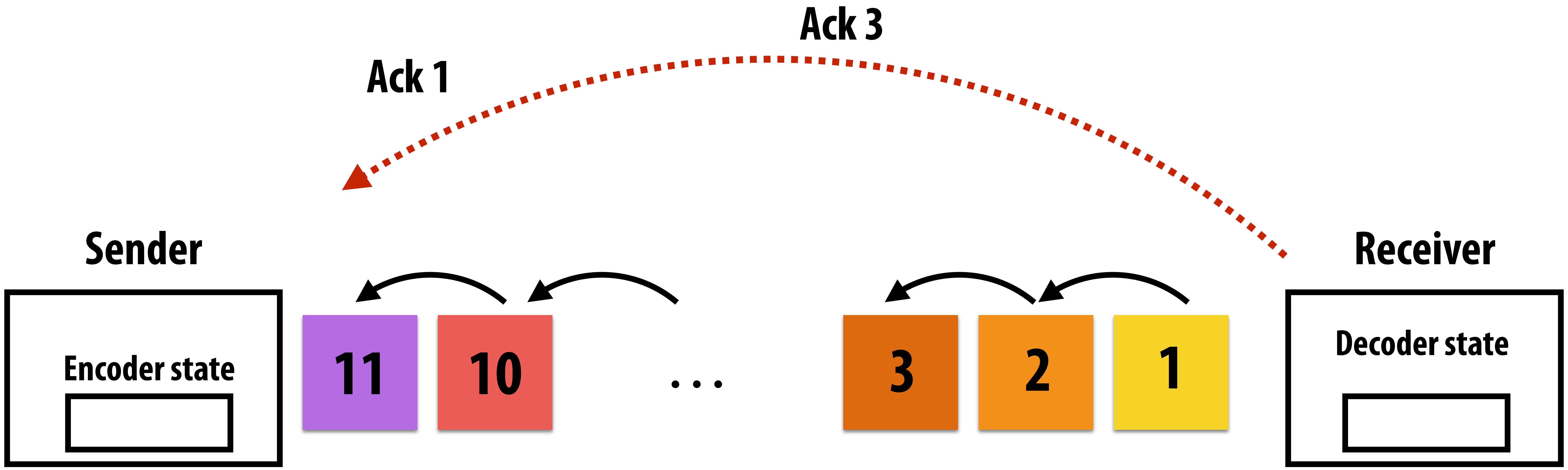
Protocol: attempts to determine and use the available capacity of the network

But generates individual frames
(which individually may or may not exceed network capacity)

- If the encoder overshoots, packet loss occurs. As a result, frames get dropped

[Credit: Fouladi et al. 2018]

Consider challenges



Sender realizes packet carrying frame 2 has been dropped (e.g, it was too big)

But sender cannot re-encode frame at lower size because it’s moved on and has different internal state

Stateless (functional) video encoder

```
// prob model: tables representing encoding of values in video stream
// reference_images contains three prior images
state:= (prob_model, reference_images[3]);

// just a full image
keyframe := image pixels for entire frame

// prediction_modes and motion vectors define how to predict current
// frame given decoder state
// residue is correction to this prediction
interframe := (prediction_modes, motion_vectors, residue)

// decoding a frame generates one image of pixels, and
// an updated decoder state
decode(state, compressed_frame) -> (new_state, image)

// generate an interframe approximating image given the current
// decoder state.  This operation requires expensive motion estimation.
encode-given-state(state, image, quality_param) -> interframe
```

[Fouladi et al. 2017]

Salsify: codec presents network three options

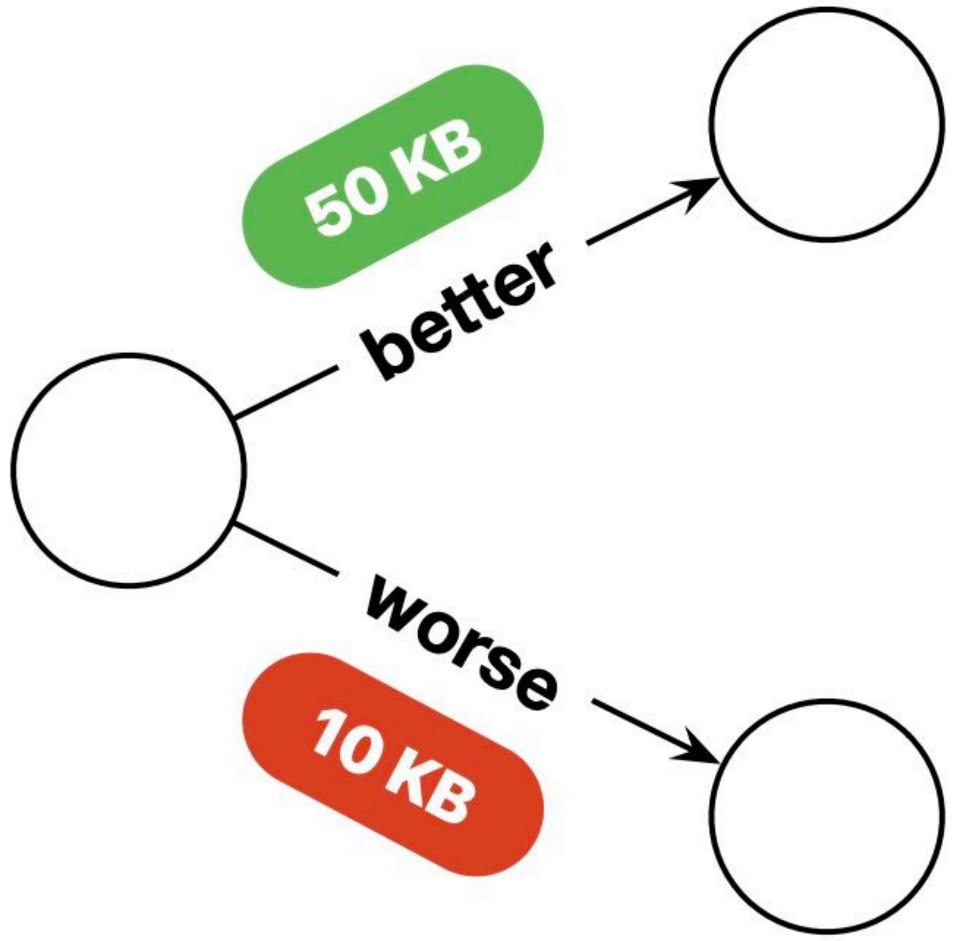
[Fouladi et al. 2018]

For each frame, codec presents the transport with *three* options:

▲ A slightly-higher-quality version,

▼ A slightly-lower-quality version,

✕ Discarding the frame.



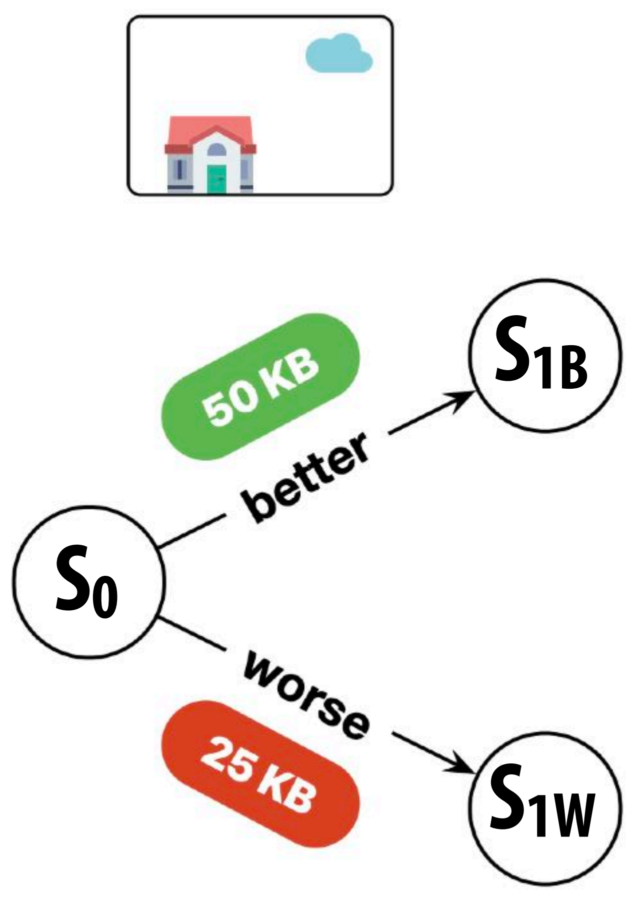
Notice roll of functional encoder.

Can encode "better", reset to previous state, and then encode "worse".

Salsify’s “video aware transport protocol: network determines what to transmit based on size of compressed frames

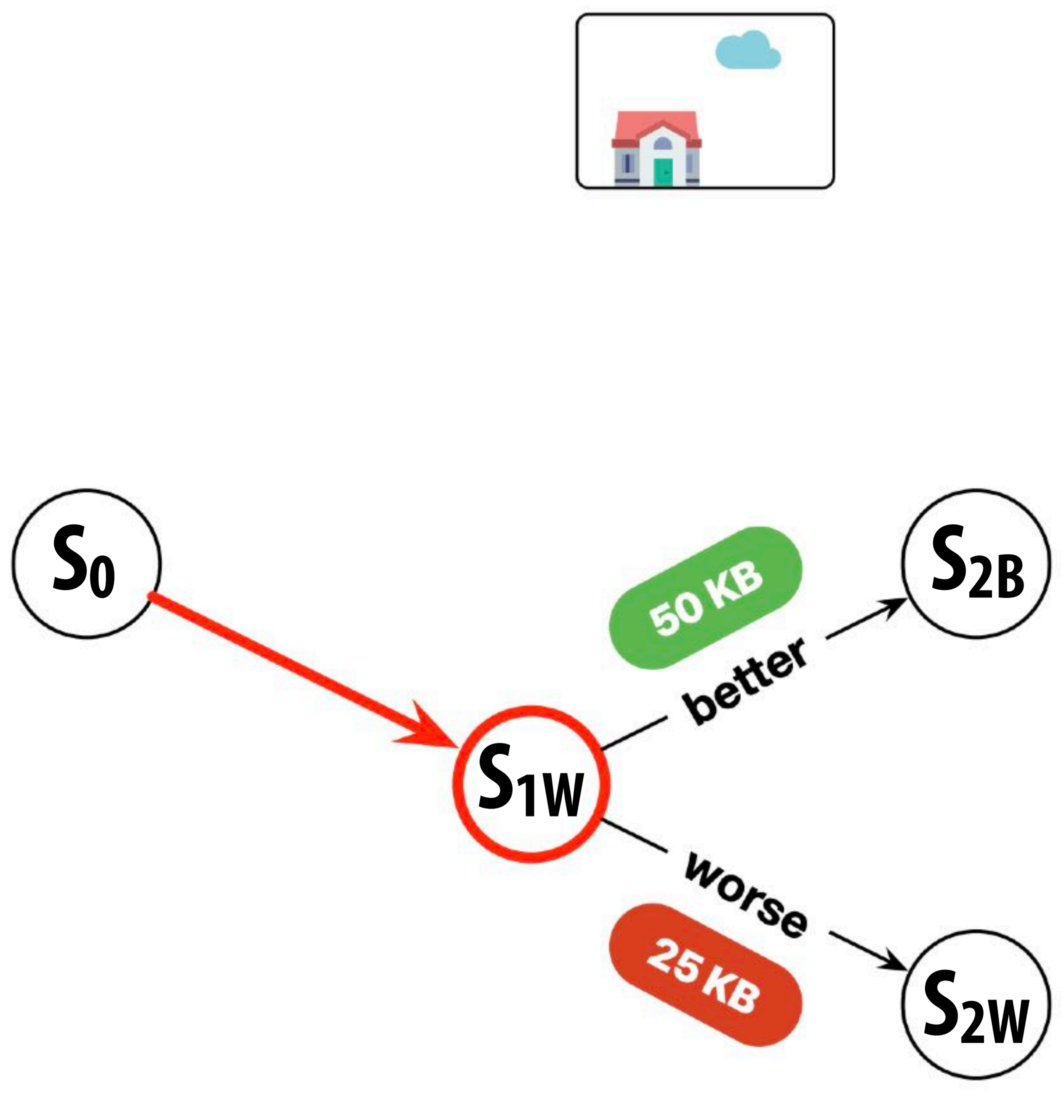
Before: network tried to send whatever the compressor generated.

Notice roll of functional encoder.
Can resume encoding from state that results from transport’s choice.



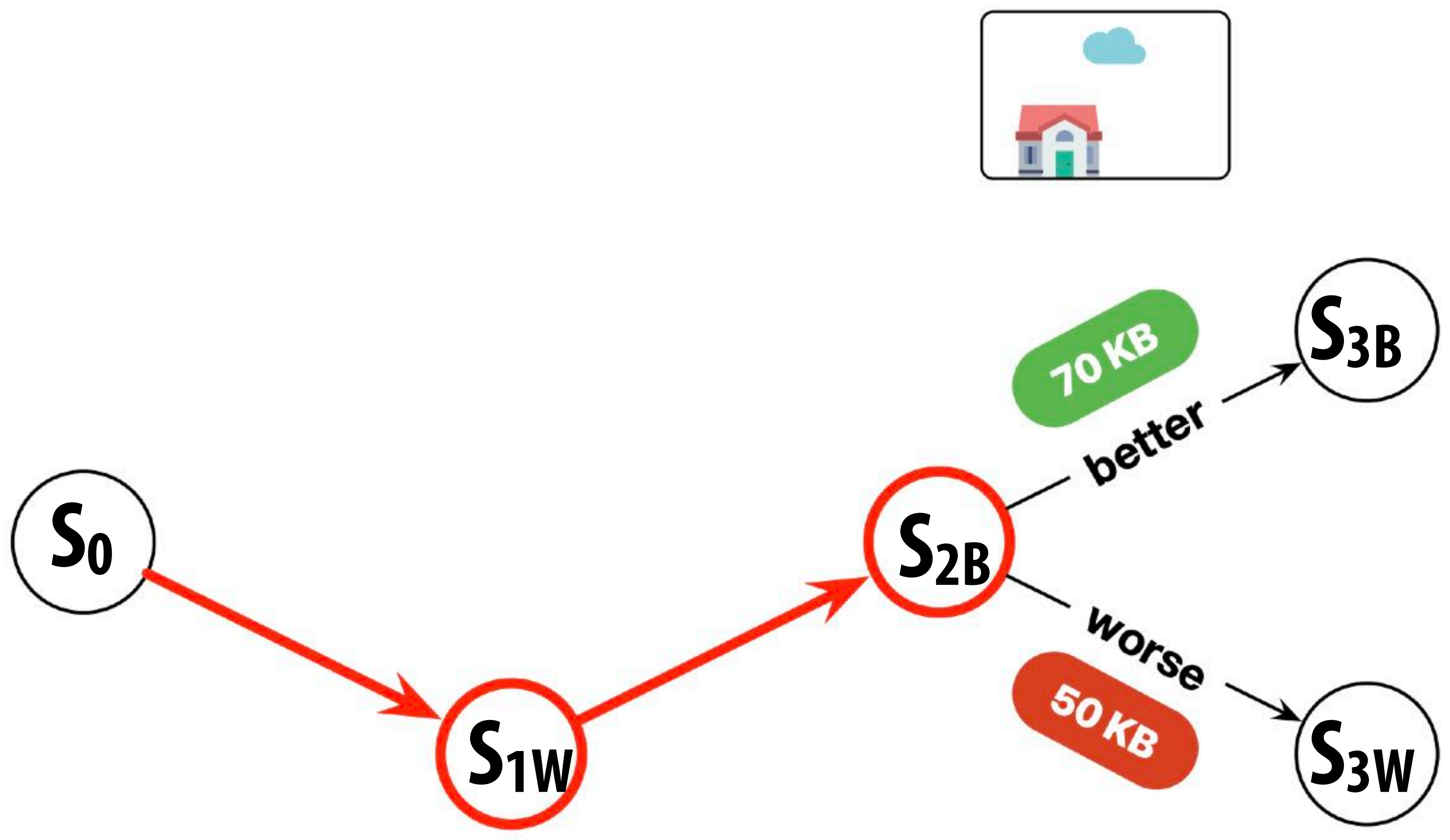
target frame size 30 KB

Transport picks “worse”



target frame size 55 KB

Transport picks “better”

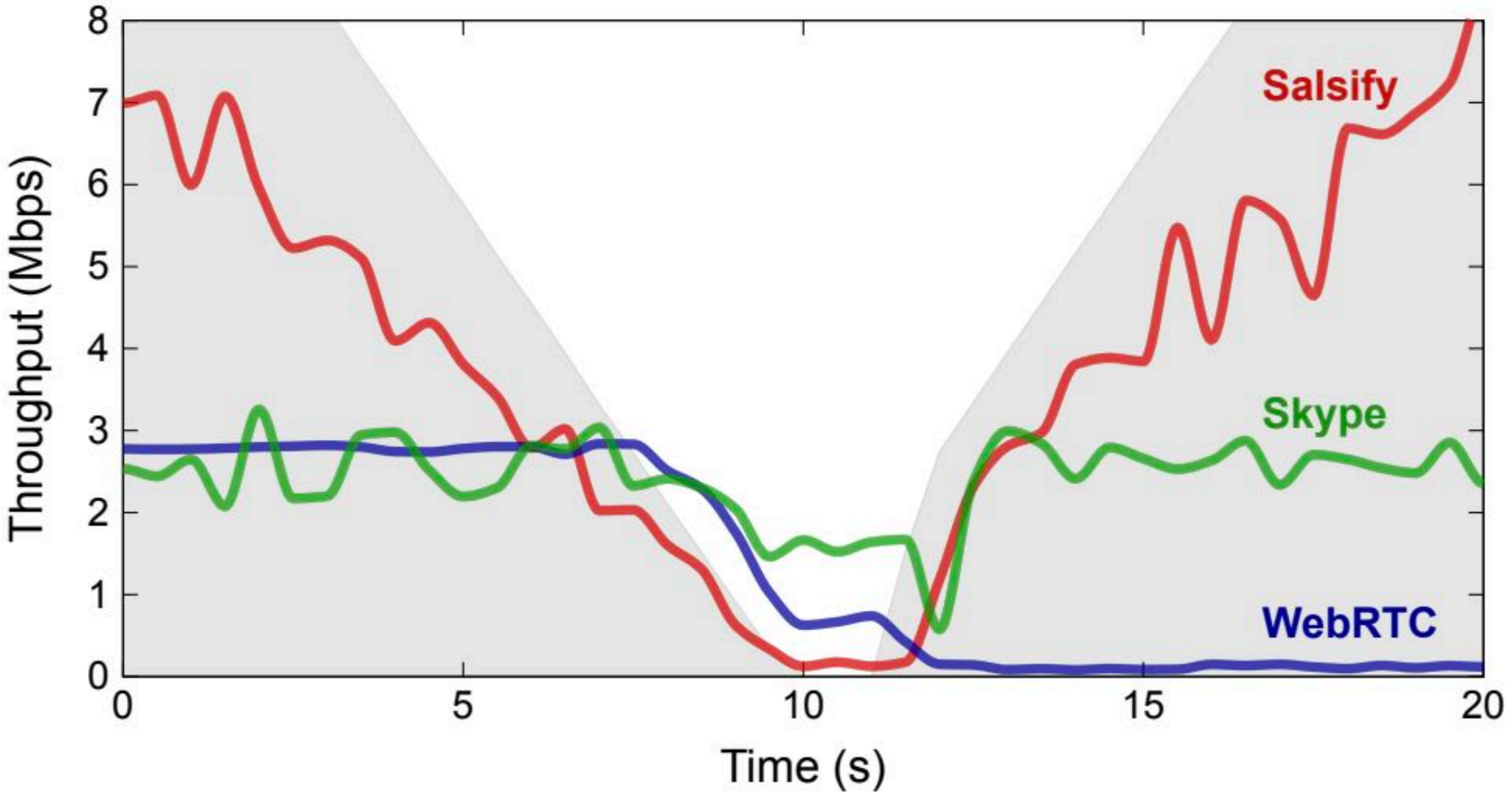


target frame size 5 KB

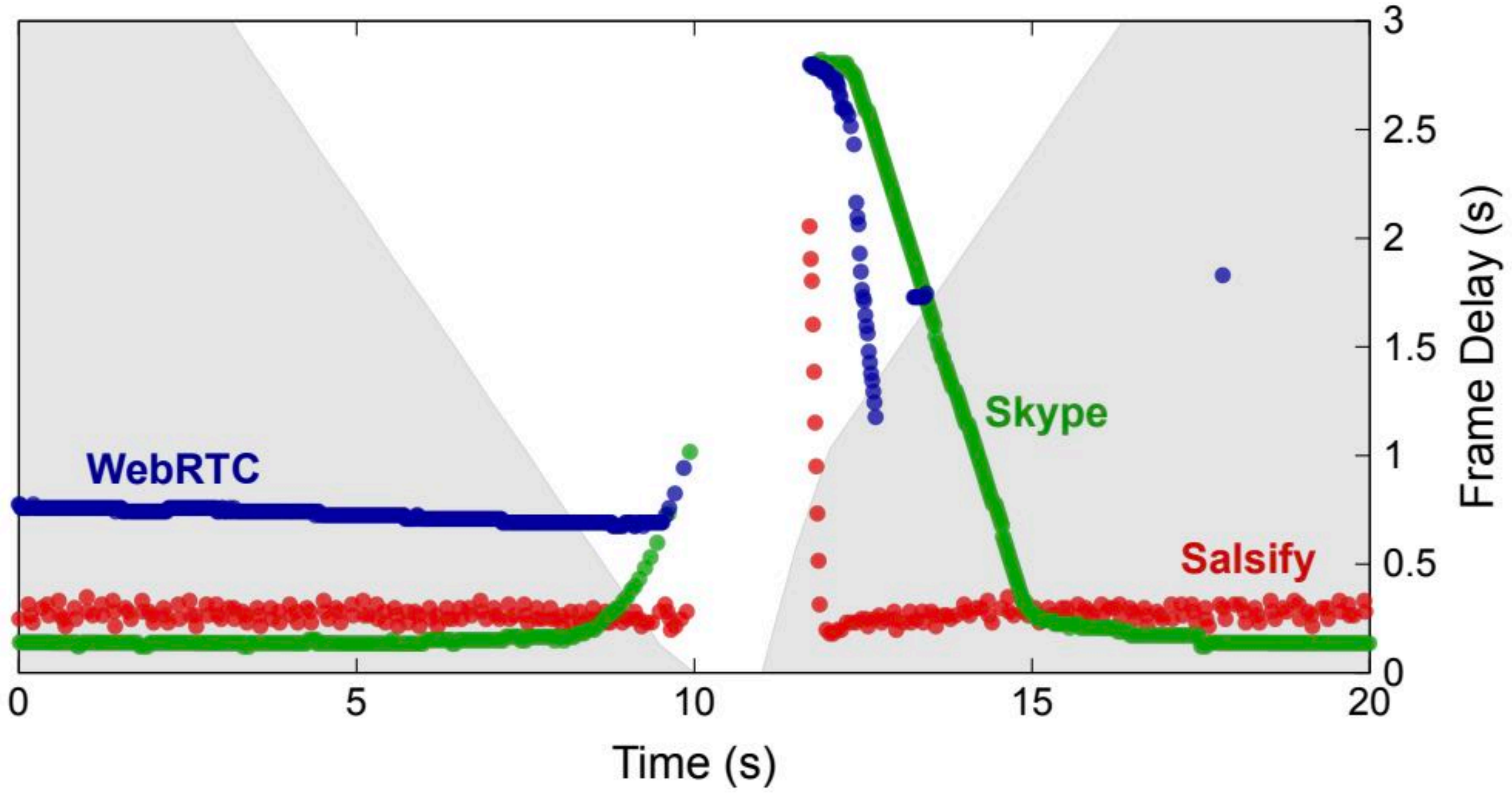
Transport picks “none”
(Encoder will proceed to compress frame 4 based on state S_{2B})

[Fouladi et al. 2018]

Much faster recovery from network changes



(a) Throughput



(b) Frame delay

Gray region shows capacity of network:
(Simulating an outage at 10 seconds)