Visual Captions

Augmenting Verbal Communication with On-the-fly Visuals

Xingyu “Bruce” Liu, Vladimir Kirilyuk, Xiuxiu Yuan,
Alex Olwal, Peggy Chi, Xiang "Anthony" Chen, Ruofei Du

github.com/google/archat

CHI23
Hamburg, Germany | Hybrid
April 23-28, 2023
reCHInnecting

Google

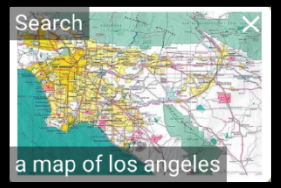
UCLA HCI RESEARCH

Systems to Facilitate Verbal Communication

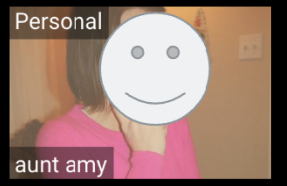
mu-kvh

Visual Augmentations of Spoken Language

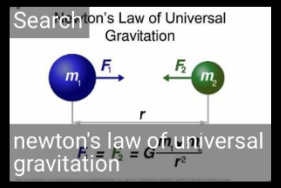
"So where do you want to visit in LA?"



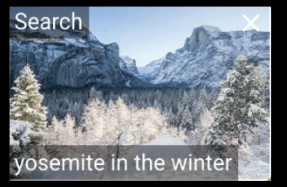
"Your aunt Amy will be visiting this Saturday."

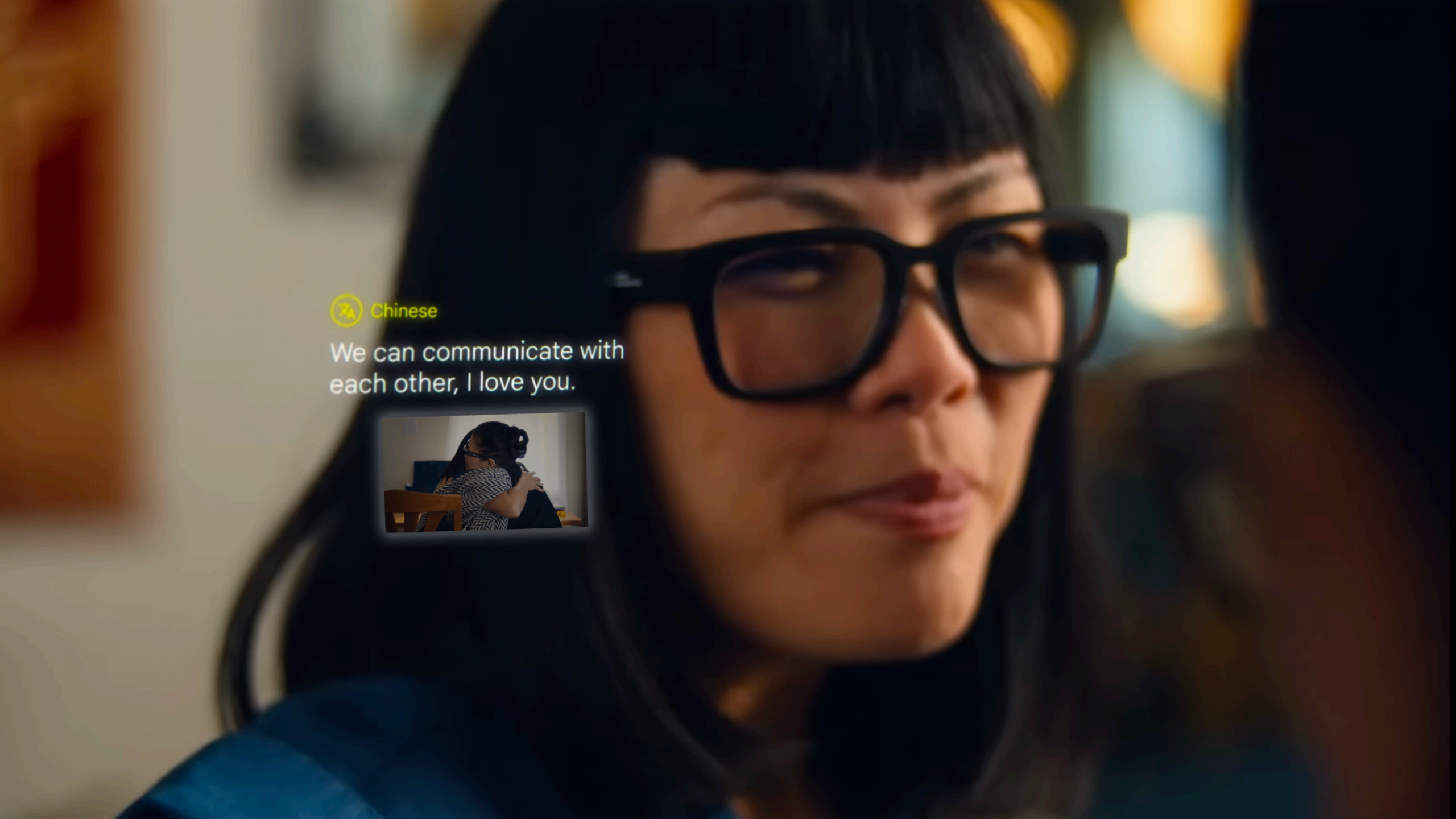


"We will cover the Newton's Law of Universal Gravitation"

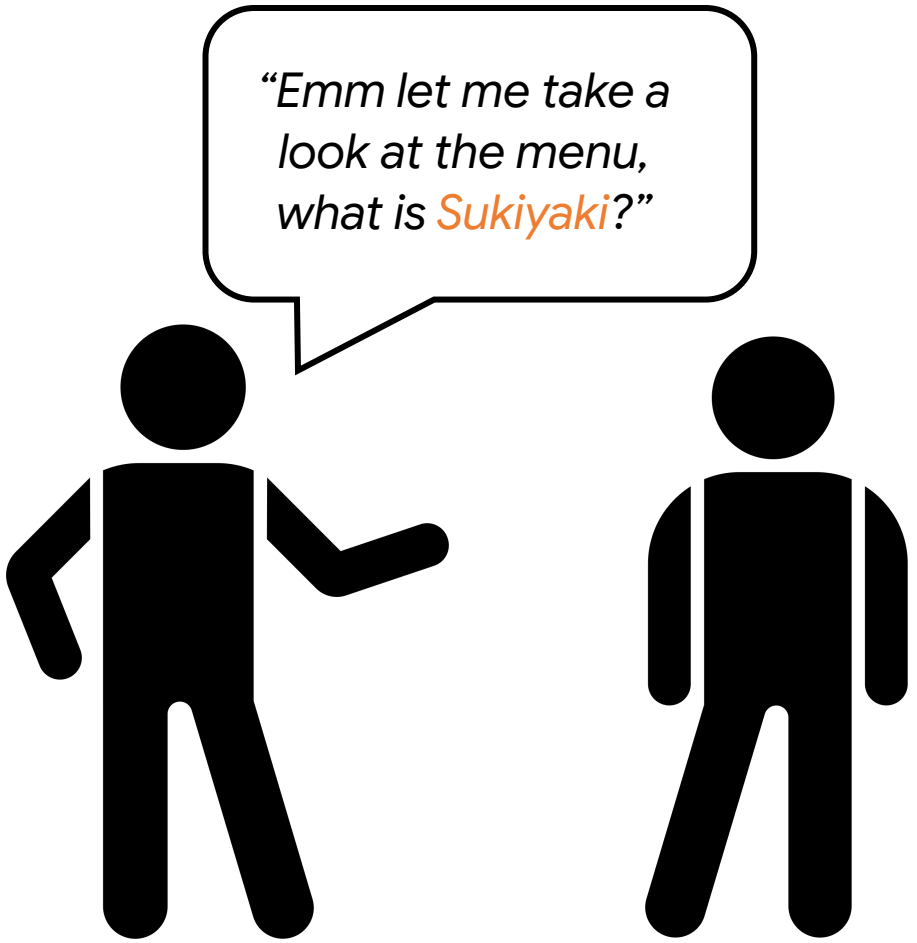


"Yosemite in the winter is really beautiful."




Chinese
We can communicate with each other, I love you.

Motivation



Motivation

Visual Captions

A. VC1.5K Dataset

"Me and my family went to Disneyland, it was so fun!"

→

<a photo> of <Disneyland> from <online image search>

<a photo> of <me and my family at Disneyland > from <personal album>

<an emoji> of <happy face> from <emoji search>

"So where do you wnat to visit in LA?"

→

<a map> of <Los Angeles> from <online image search>

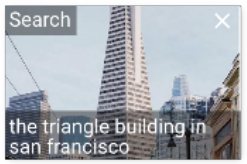
...

B. Visual Prediction Model

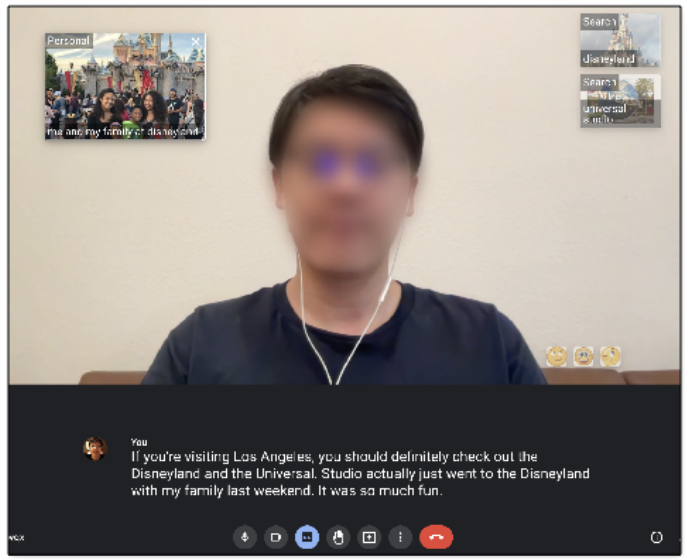
"Tokyo is located in the Kanto region of Japna"

"We spent our weekend in Yosemite"

"You know, the triangle building in San Francisco."



C. Visual Captions Interface



Formative Study (N=10)

Equation and 3D curve/surface

When explaining scientific idea, show equations or interactive plot, e.g. sin curve, or 3D vis of a mesh that allows deformation.

Demo video

Do real-time generation of demo video based on verbal description, e.g. steps of making a cake.

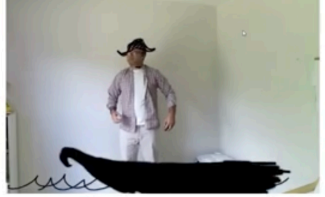
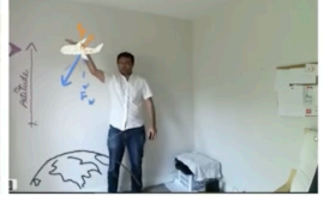
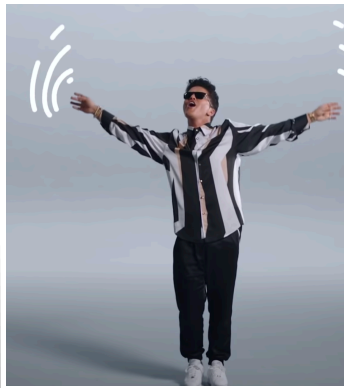
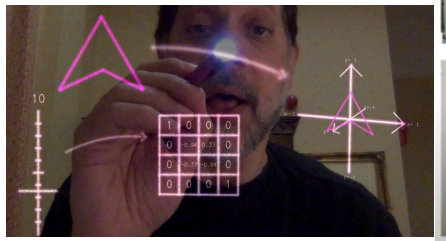
Presentations
Bring presentation concepts to life for lighting talks etc

Classroom
Limit search space to niche topic area eg WW2 for history lesson as teacher speaks - limit to just photo / muted video loops etc.

Visual mind map search
Have idea of something in mind but unsure how to express - work with AI to find right imagery

Predictive Utility
Combine with topic prompts to visualise what to speak about next - predictive to give hints to presenter what to touch on next



Shopping
Show examples of outfit or similar designs or material/wash information etc.

Driving
Show street, signs, pedestrian stats.

Indoor navigation
Show objects information.

ADHD
For people who can't tell other people's emotion, show their emotions.

Hearing impaired
Show sounds around them in images.

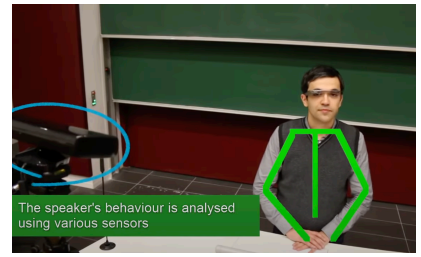
Education
A math teacher demonstrate a orientational relationship between different objects.

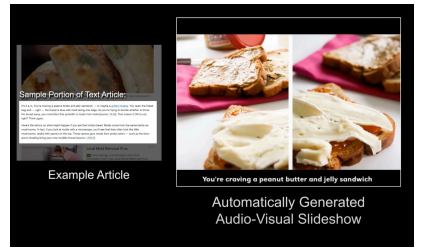
Casual talk with friends
Talk about the recent movie (shows the movie posters); books (with information related to the author / review about the book)

Describe dimensions
Talk to your friends about the size of the object -> visualize it in real world

Visual notes?
Discuss the topic with the user, save the text + pictures into a summary -> may send to the user's phone for memory purpose

Collaboration
Demonstrates a 3D concept to your friends, make modifications based on the discussion (either by gesture, or language)



Visualizing go/links
When people are mentioning a go link in the company, show a preview of the slidesdeck / doc and people may click into it.

When people are mentioning a website like World Wide Fund for Nature, show its website

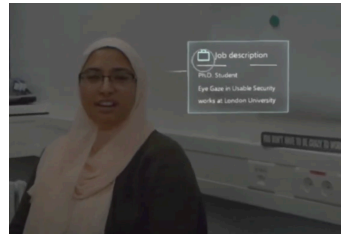
Visualizing people names
From your contacts photos

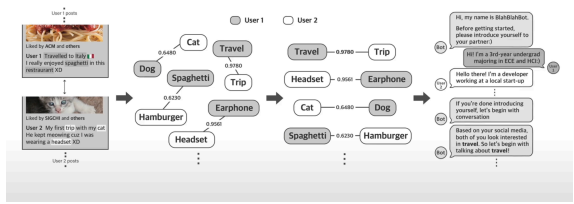
Alternative sayings
Replace basic words with advanced words -

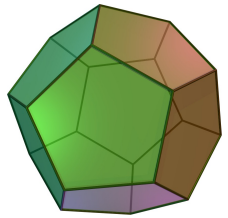
Japan culture - hint users to use polite

When saying something, hint "please"

Visualize balance
When talking about shopping something, show price of your balance; when talking about eating something, show caloryx of the food







Risk: Misleading
The audience could capture wrong ideas and ponder / get distracted due to irrelevant images (e.g., that "United" airplane image in this session)

Multi-language conversation?
For people communicating in different language, using pictures as middle bridge?

E.g., wait you at building PR-55

Eye-candy experience
Celebration

Quick Select/Remove
Swipe away if something goes wrong

Resize & Animation
Based on gestures (pinch, rotate, etc.), zoom to see more related images for quick selection

Pin
Pin an image & avoid overwriting by other images



Presentations
Bring presentation concepts to life for lighting talks etc

Classroom
Limit search space to niche topic area eg WW2 for history lesson as teacher speaks - limit to just photo / muted video loops etc.

Visual mind map search
Have idea of something in mind but unsure how to express - work with AI to find right imagery

Predictive Utility
Combine with topic prompts to visualise what to speak about next - predictive to give hints to presenter what to touch on next

Elevator Pitch
Meeting with Shahram in elevator, pitch latest project ideas to Shahram by retrieving recent Slides from gDrive when mentioning project code names.

Hometown Introduction
In business meetings, when meeting with people throughout the world, introduce hometown, favorite restaurant, pull them on a map in AR and directly show people where they are

Introducing Pets
When talking about pets with other people, pull the favorite dog / cat picture from gPhotos for introduction

Visualizing dishes
There are many unknown dishes in Japanese menus, cameras are not always available on gGlasses. E.g., talk about Sushi, Sashimi, Donburi, Udon, Tempura

Visualizing math / physics
When talking about an abstract concept such as Euler's

formula, Newton's law of universal gravitation, visualize them

When talking about numbers, equations for grocery, do the compute for the users.

Privacy filter
Mosaic unwanted pictures if needed

Children mode
Only use selected safe pictures

Cultural difference
Changes the visualization depending on the region/location
purse , biscuits

ASL + speech

Show ASLs.

BCI
Visualizing brain signals

Eye tracking + Gaze
Use eye tracking to navigate between various visuals and use gaze to select one

Pros: intuitive, no learning curve

Cons: eye fatigue
- Error-prone target selection due to people unconsciously blinking their eyes
Frastrusion

Hand Gesture
Utilizing mid-air gestural input to activate/deactivate and shuffle between visuals

Pros: direct input
No extra mappings needs to take

Cons: muscle fatigue for continuous usage

Head Gesture
Present visuals up to 3 once a time. Users can shake their head to confirm/select/activate the left, middle and right visuals

Pros: somehow intuitive
Users can proceed with confidence

Cons: fatigue in the neck

Voice Input
Interpreting the human voice and generate command based on that.
People could say "choose this one" or "I like the visuals on the top right corner".

Pros: straight forward

Cons: human languages are vague and may be difficult to generate precise commands
Frastrusion

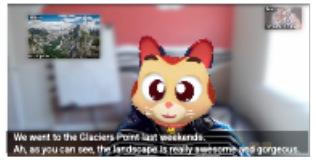
Design Space

Generating Visuals

1 Temporal

Synchronous

Real-time: Generating visuals on-the-fly



Asynchronous

Pre-edit: Configure visuals beforehand

Related Work: Body Driven Graphic [42]

Post-hoc: Edit visuals afterwards

Related Work: CrossCast [52]
Audio Visual Slideshows [24]

2 Subject

Express

Visualize **one's own** speech

Understand

Visualize **others'** speech

3 Visual

Visual Content

Generic

Contextual

Visual Type — *Abstract* ⟷ *Concrete*

Visual Effects

Emoji/Symbol

Image

Gifs/Video

3D Model

Visual Source — *Personal* ⟷ *Public*

Personal

Social Media

Knowledge Base

Online Search

Interaction with Visuals

4 Scale

One-to-one

One-to-many

Many-to-many

5 Space

Co-located

Remote

6 Privacy

Private

Shared

Public

7 Initiation

Human-controlled ⟷ *Fully Automatic*

On-demand-suggest

Auto-suggest

Auto-display

8 Interaction

Speech

Gesture

Body

Facial Expression

Gaze

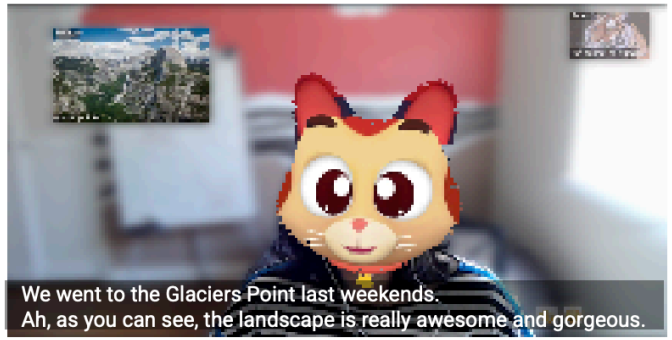
Input Device

Design Space

1 Temporal

Synchronous

Real-time: Generating visuals on-the-fly



Asynchronous

Pre-edit: Configure visuals beforehand

Related Work: Body Driven Graphic [42]

Post-hoc: Edit visuals afterwards

Related Work: CrossCast [52]
Audio Visual Slideshows [24]

King Kamehameha II marked a new era by ending traditional practices in 1819.

Design Space

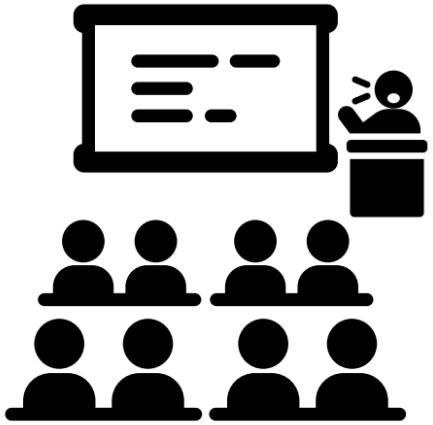
2 Subject

Express

Visualize **one's own** speech

Understand

Visualize **others'** speech



Design Space

3 Visual

Visual Content

Generic

Contextual

Visual Type

Abstract ⟷ ***Concrete***

Visual Effects

Emoji/Symbol

Image

Gifs/Video

3D Model

Visual Source

Personal ⟷ ***Public***

Personal

Social Media

Knowledge Base

Online Search

Design Space

Interaction with Visuals

4 Scale

One-to-one

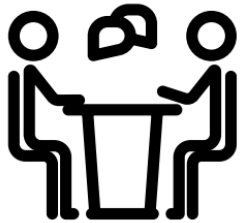
One-to-many

Many-to-many

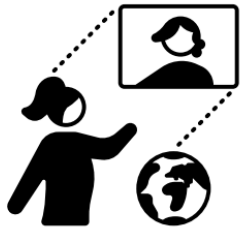
Design Space

5 Space

Co-located



Remote



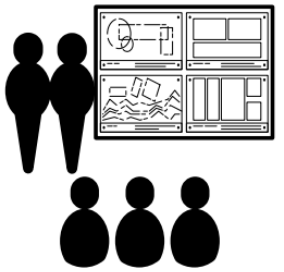
Design Space

6 Privacy

Private

Shared

Public



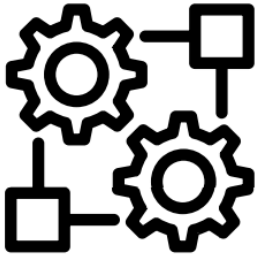
AI Proactivity

7 Initiation

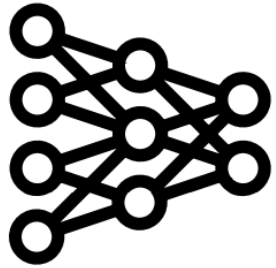
Human-controlled ⟷ ***Fully Automatic***

On-demand-suggest

Auto-suggest



Auto-display



Design Space

8 Interaction

Speech

Gesture

Body

Facial Expression

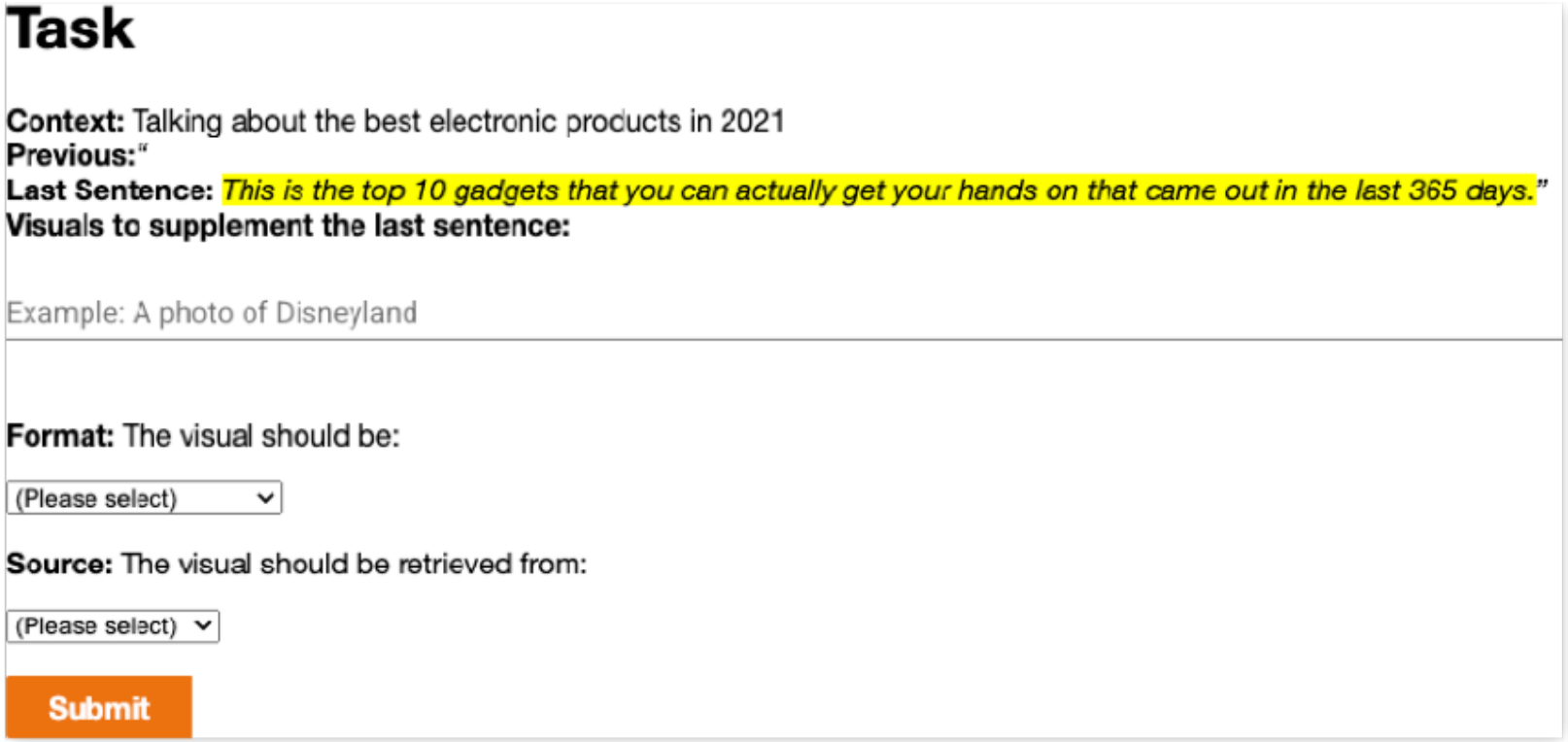
Gaze

Input Device

VC1.5K Dataset

1595 sentence-visual pairs from 42 YouTube videos and the Daily Dialog datasets

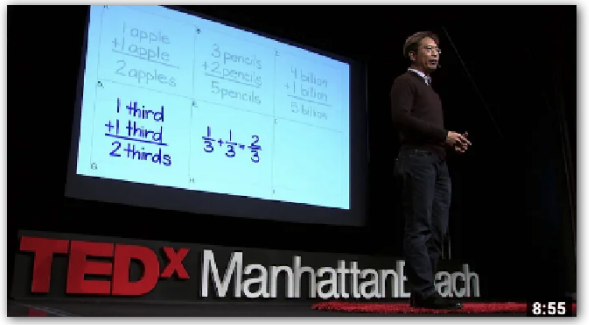
246 MTurk workers



VC 1.5K


4:20
8:01


14:39


1 third
+1 third
2 thirds
TEDx Manhattant ach
8:55


A.V. CLUB
5:48

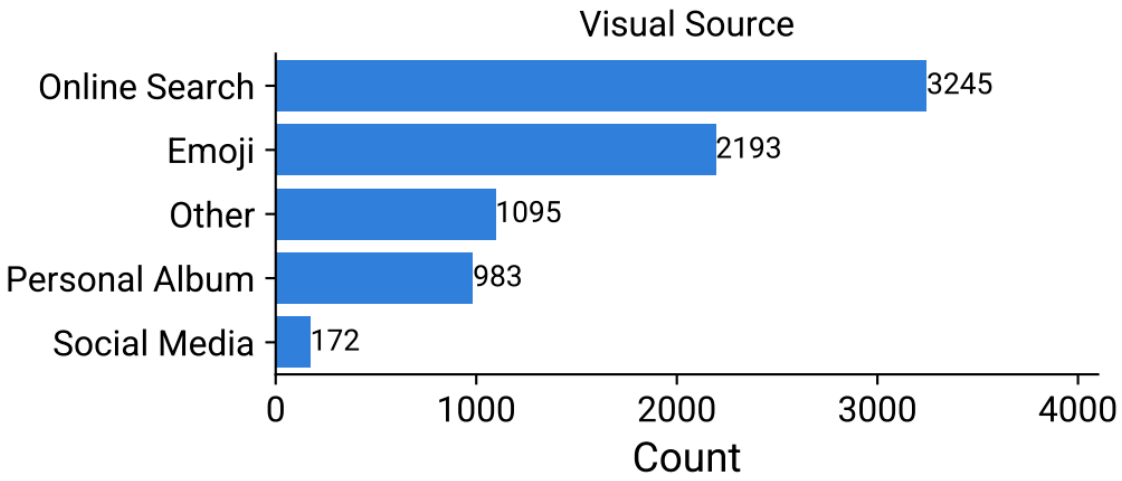

48:05

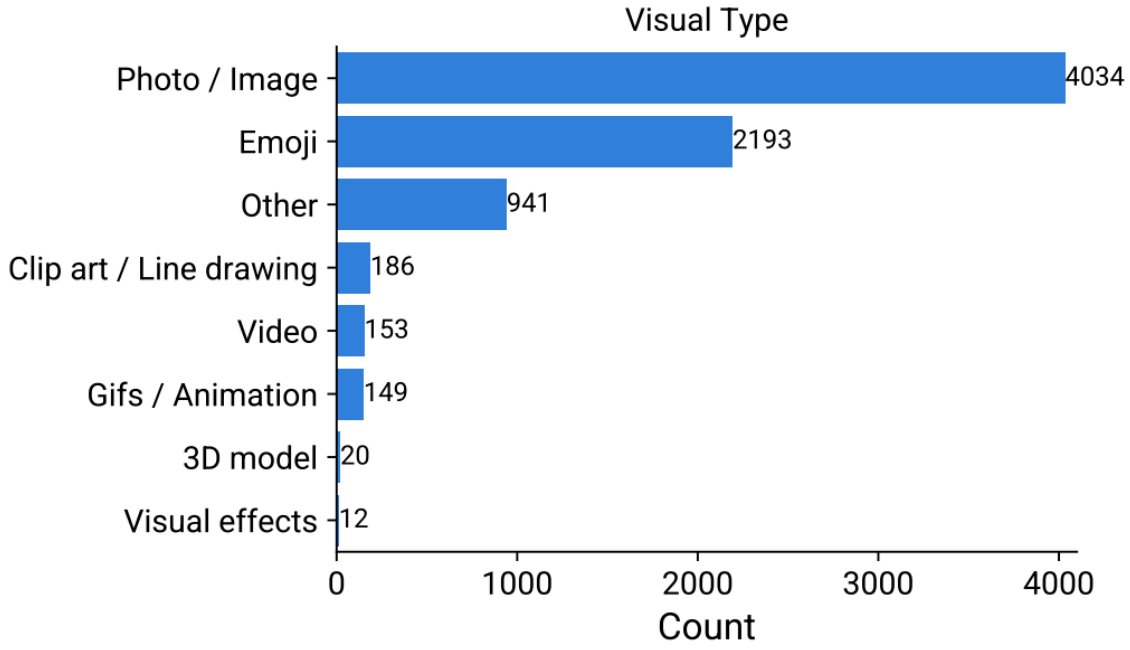

TRAVEL GUIDE
Los Angeles
8:50


FULL EPISODE
HISTORY
42:35


No, thank you.
26:55

VC 1.5K

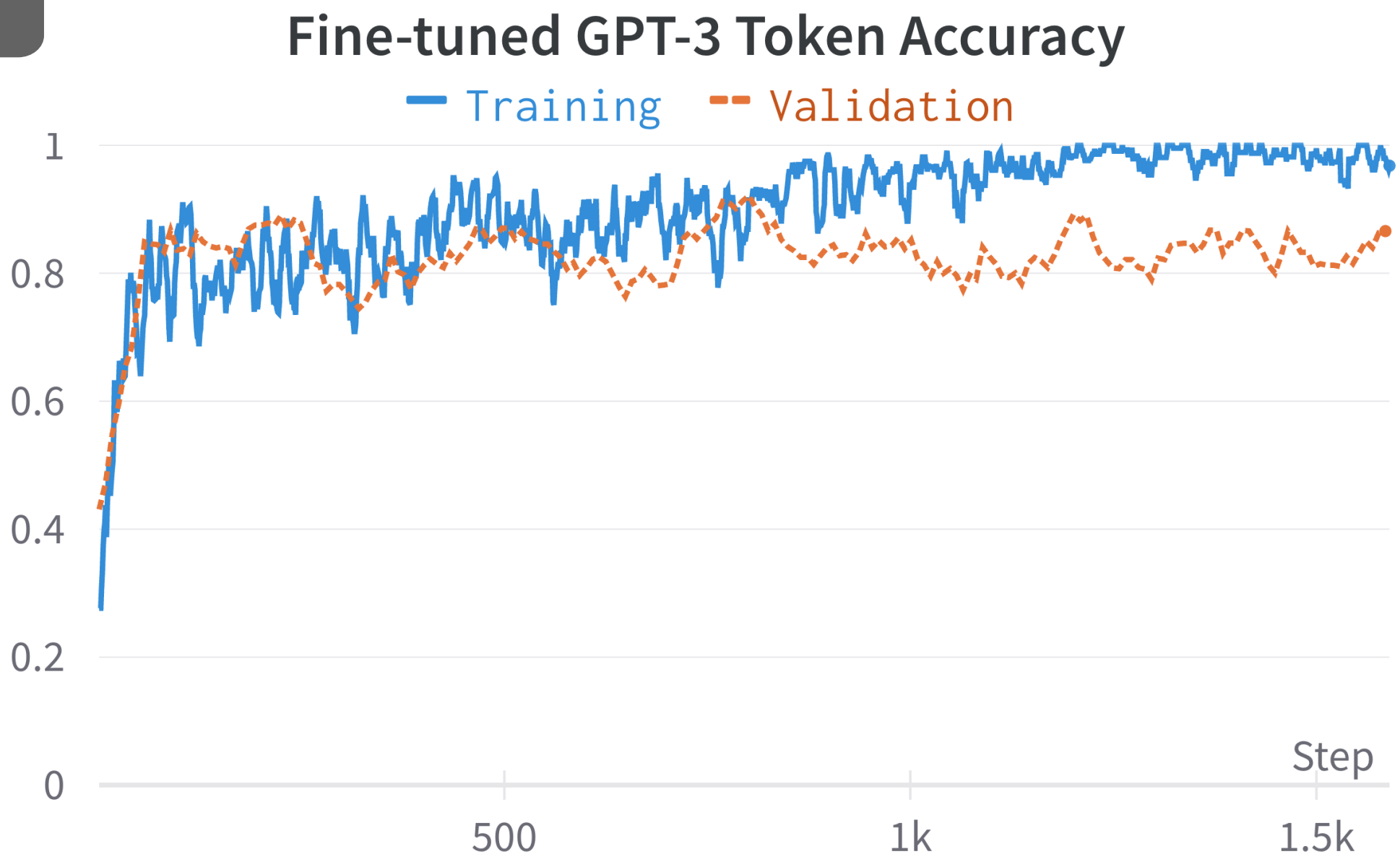

Visual Source
Online Search
Emoji
Other
Personal Album
Social Media
3245
2193
1095
983
172
0
1000
2000
3000
4000
Count


Visual Type
Photo / Image
Emoji
Other
Clip art / Line drawing
Video
Gifs / Animation
3D model
Visual effects
4034
2193
941
186
153
149
20
12
0
1000
2000
3000
4000
Count

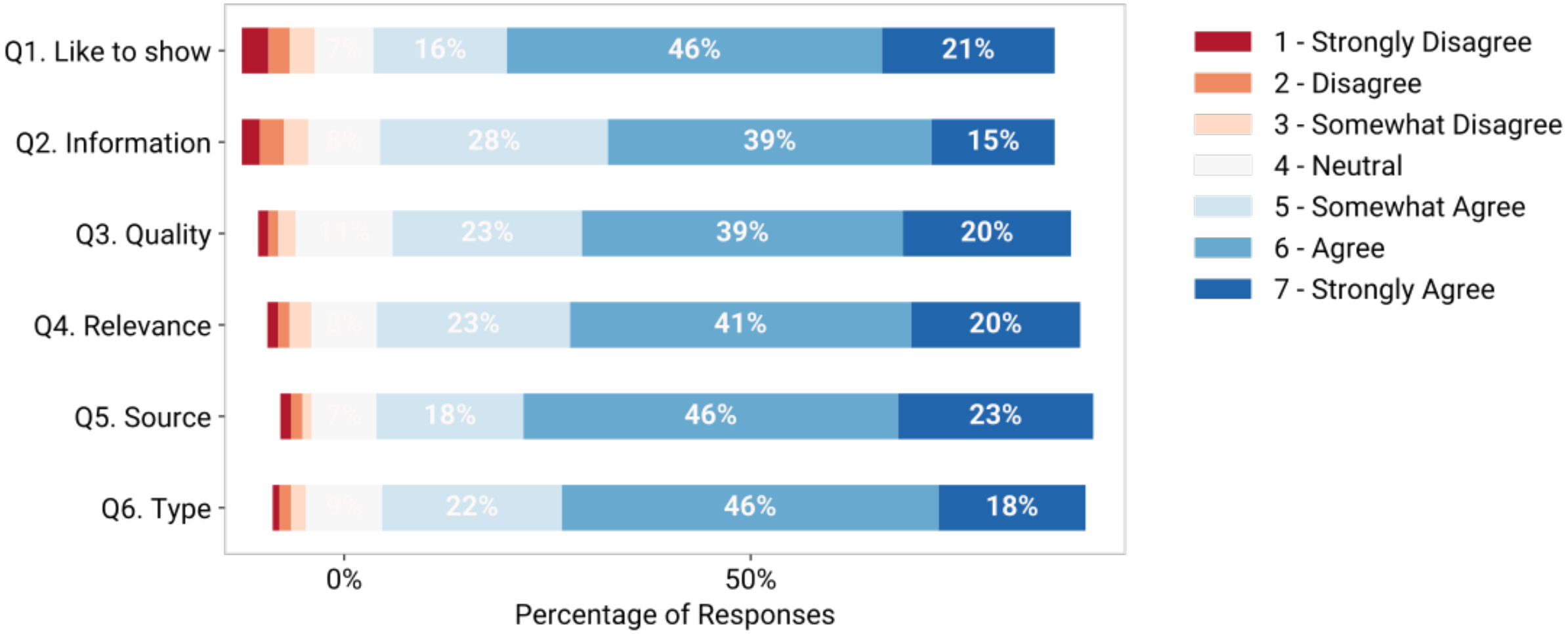
Model Prompt

{“prompt”: “<Previous Conversation> →”,
“completion”:
“<Visual Type 1> of <Visual Content 1> from <Visual Source 1>;
<Visual Type 2> of <Visual Content 2> from <Visual Source 2>;
...
\n”}

86% Token Accuracy



Crowdsourced Evaluation
846 Tasks



A. Transcription Parsing

"I just went to Disneyland with my family last weekend, it was super fun! And..."

Speech-to-Text

100 ms interval

"I just went to Disneyland with my family last weekend, it was super fun!"

Retrieve last complete sentence and last sentence > n_min

Fine-tuned Language Model

B. Visual Intents Prediction

<A photo> *of* ***<Disneyland>*** *from* ***<online image search>***

<A photo> *of* ***<me and my family at Disneyland last weekend>*** *from* ***<personal album>***

<A emoji> *of* ***<happy face>*** *from* ***<emoji search>***

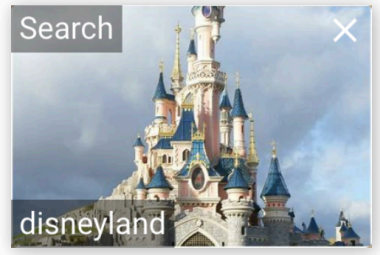
Predict Visual Intents

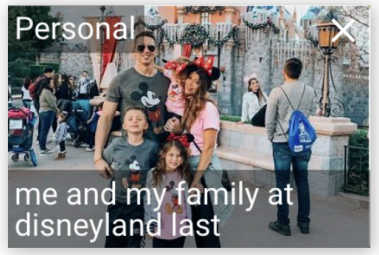
C. Visual Generation

Online Image Search

Text Image Retrieval

Custom Database





Retrieve Relevant Visuals

ARChat
Chrome Plugin

You

2:03 PM | uys-pdyp-bie

AI Proactivity

Auto Display

AI Proactivity

On-demand Suggest

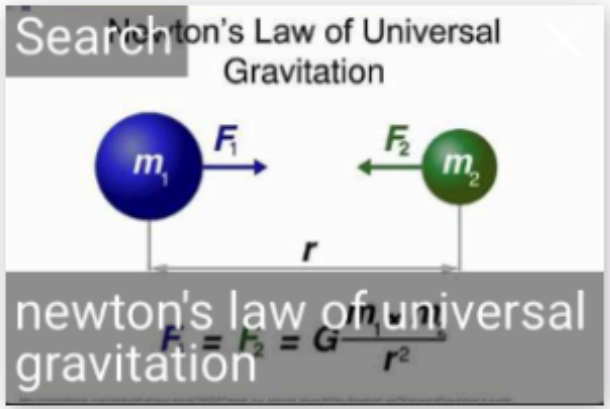
AI Proactivity

Auto Suggest

Open Vocabulary

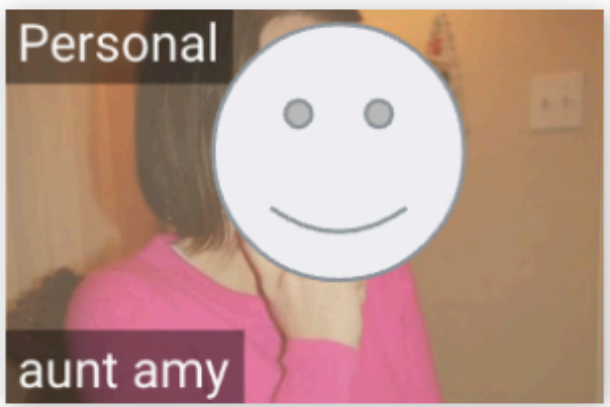
"We will cover the Newton's Law of Universal Gravitation"

(1) → Visual Content: Law of universal gravitation
Visual Type: Diagram
Visual Source: Internet Search



"Your aunt Amy will be visiting this Saturday."

(2) → Visual Content: Aunt Amy
Visual Type: Photo
Visual Source: Personal Album



Multiple Visual Suggestions

"Tokyo is in the Kanto region of Japan."

(3) → Visual Content: Tokyo
Visual Type: Photo
Visual Source: Internet Search

(4) → Visual Content: Kanto Region of Japan
Visual Type: Map
Visual Source: Internet Search

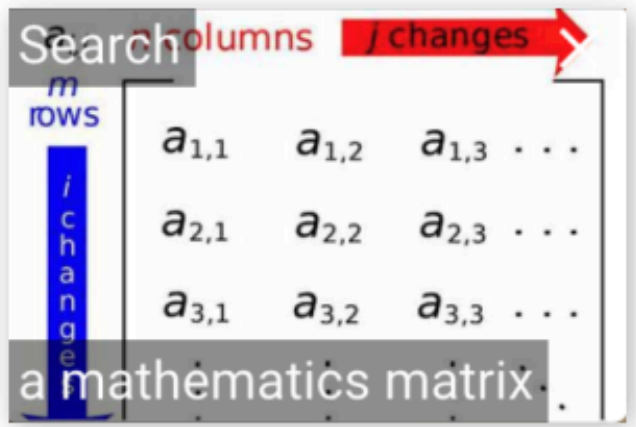
Different Visual Content

"My favorite movie is the Matrix."

(5) → Visual Content: The movie Matrix
Visual Type: Poster
Visual Source: Internet Search

"In today's lecture, we will learn a mathematical concept, matrix"

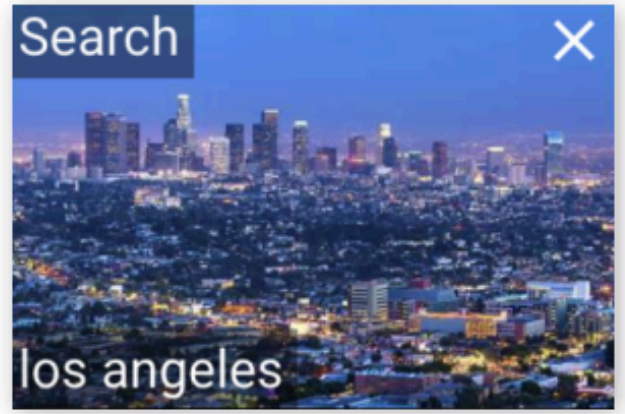
(6) → Visual Content: A math matrix
Visual Type: Diagram
Visual Source: Internet Search



Different Visual Types

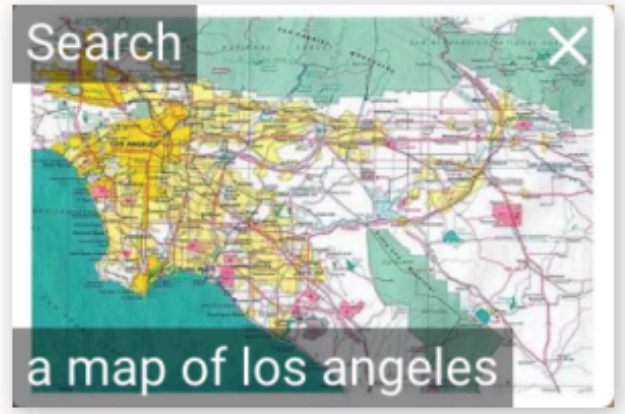
"Welcome to Los Angeles!"

(9) → Visual Content: Los Angeles
Visual Type: Photo
Visual Source: Internet Search



"So where do you want to visit in LA?"

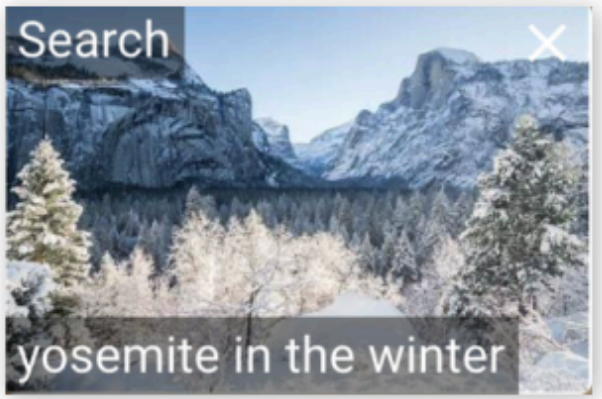
(10) → Visual Content: Los Angeles
Visual Type: Map
Visual Source: Internet Search



Different Visual Sources

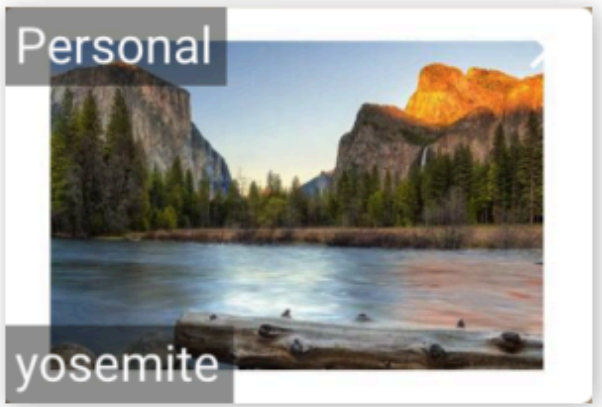
"Yosemite in the winter is really beautiful."

(7) → Visual Content: Yosemite in Winter
Visual Type: Photo
Visual Source: Internet Search

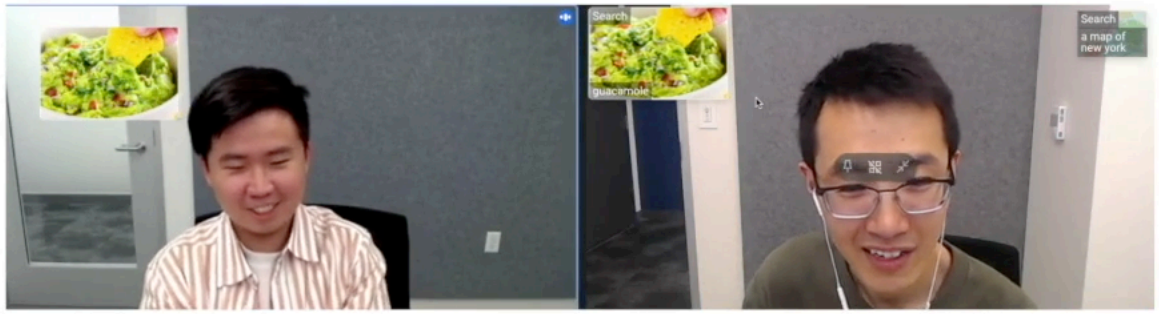


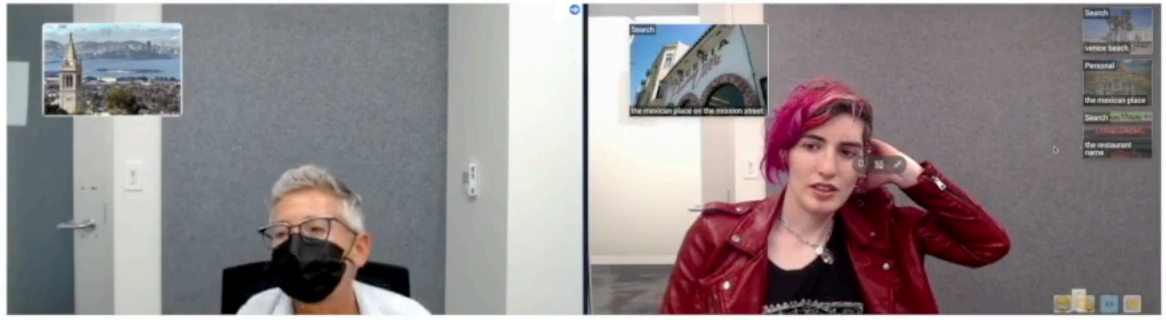
"We spent our weekend in Yosemite."

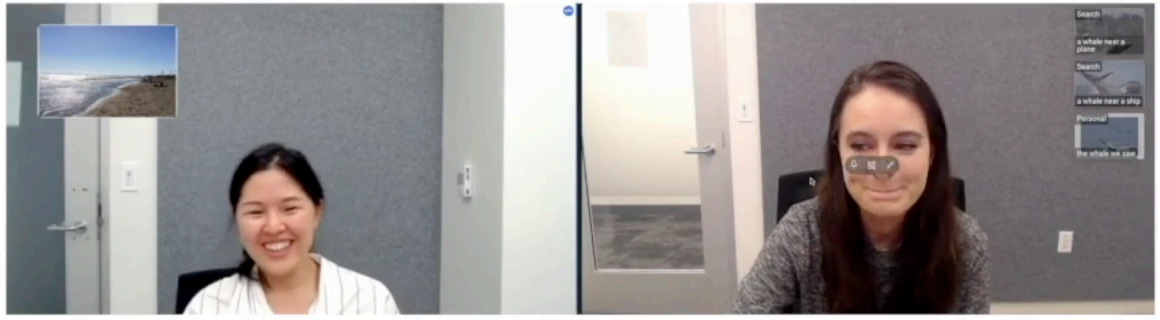
(8) → Visual Content: Yosemite
Visual Type: Photo
Visual Source: Personal Album

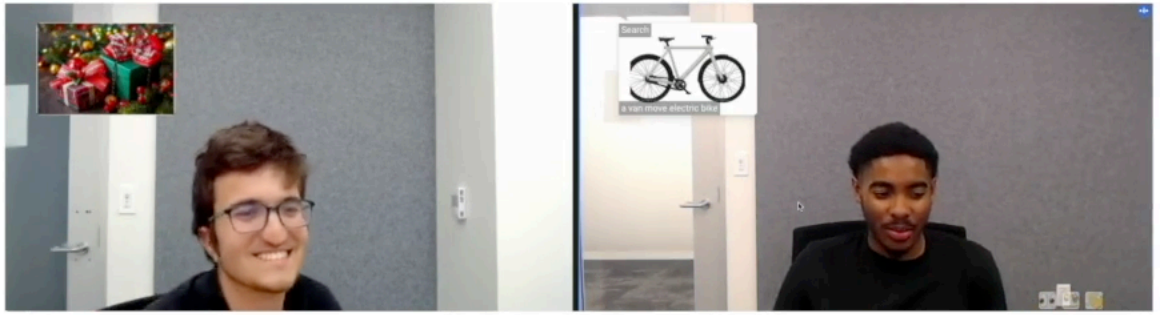


# User Studies
N=26

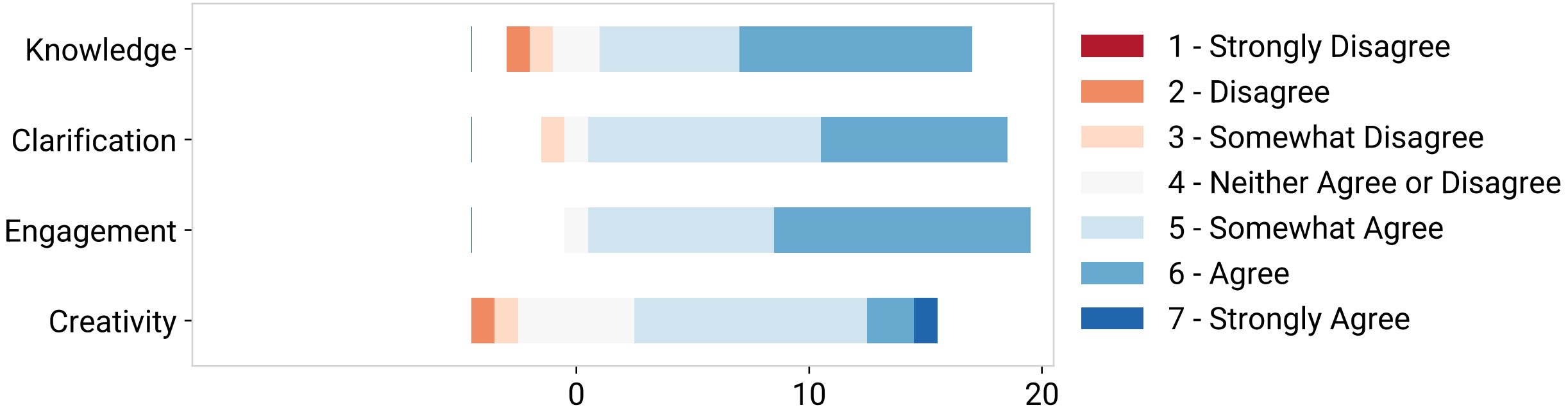








Findings


Knowledge
Clarification
Engagement
Creativity
0
10
20
1 - Strongly Disagree
2 - Disagree
3 - Somewhat Disagree
4 - Neither Agree or Disagree
5 - Somewhat Agree
6 - Agree
7 - Strongly Agree

“When I would really want visuals is like people don’t know what I was talking about. For example when I just mentioned Santa Monica Pier, it’s great that I can easily explain what it is.”

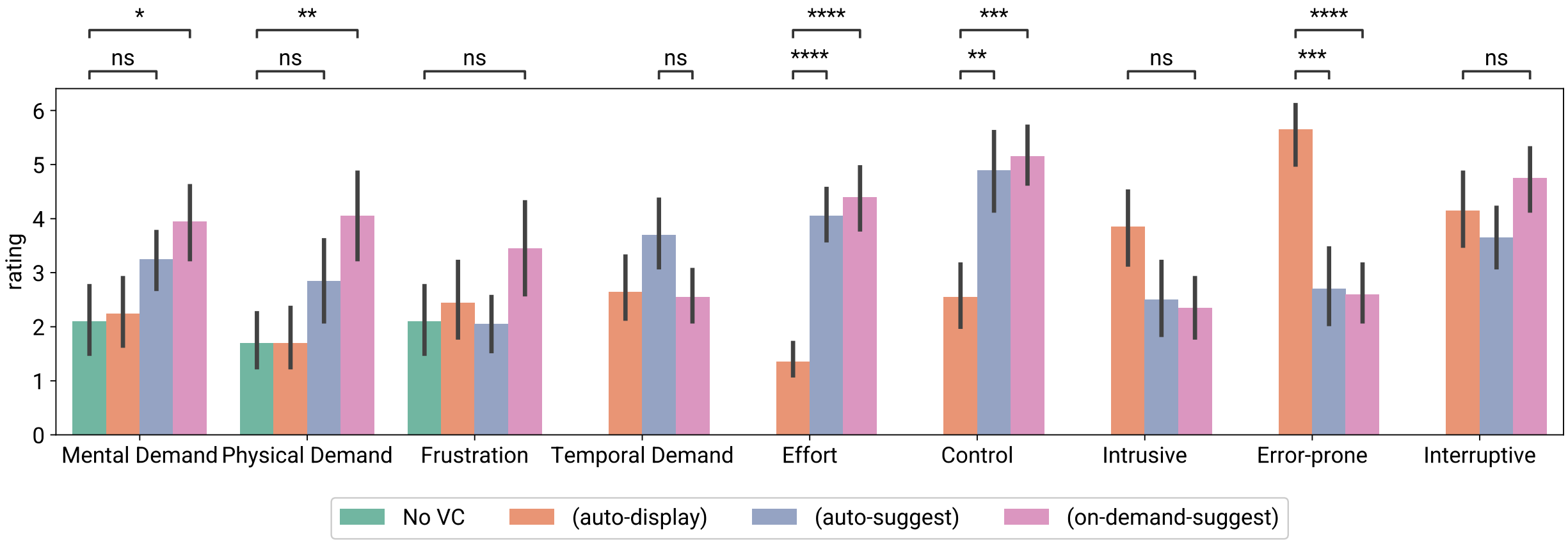
Back in the beginning when we were talking about the Avatar, there are like four or five different versions we might be discussing, the picture helped crystallize it instantly

“

It makes the conversation longer and more interactive.

”

Diverged AI Proactivity Levels



Diverged AI Proactivity Levels

*“**Not having to click** is huge for me.”*

– P7 [Auto-Display]

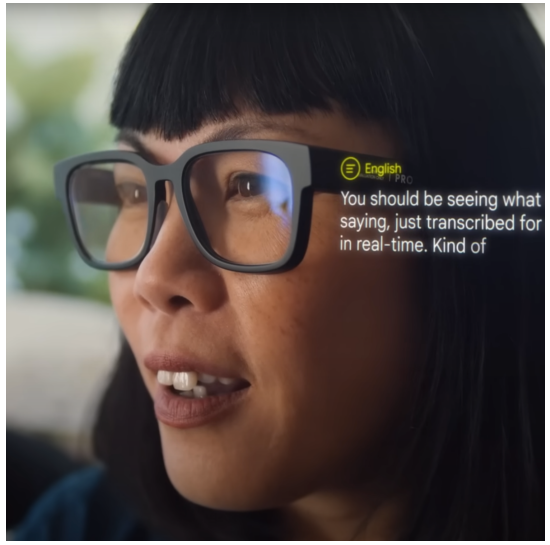
*“I like when these things pop up, so **I really know what it is like**.”*

– P8 [Auto-Suggest]

*“It’s less mental overload and distraction because I would **only activate it when I want.”***

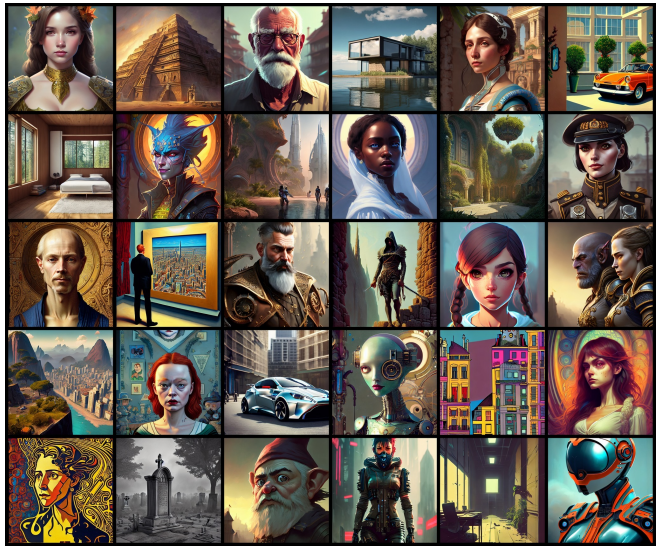
– P13 [On-Demand]

Future Work
Augmented Communication



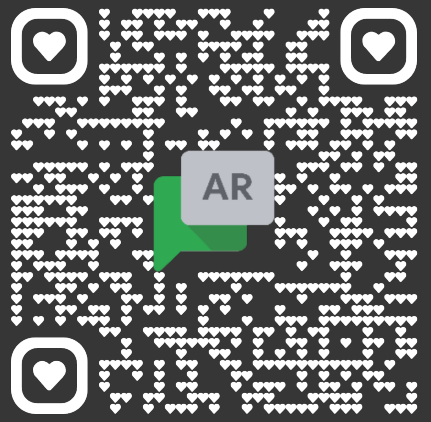
Visual Captions for in-person conversations

Personalized Visual Suggestions



Integrating text-to-image models

ARChat



github.com/google/archat

With ARChat, the CHI community can make communication more interactive, effective, and accessible with real world impact.

Visual Captions

Augmenting Verbal Communication with On-the-fly Visuals

Xingyu “Bruce” Liu, Vladimir Kirilyuk, Xiuxiu Yuan,
Alex Olwal, Peggy Chi, Xiang "Anthony" Chen, Ruofei Du

github.com/google/archat

CHI23
Hamburg, Germany | Hybrid
April 23-28, 2023
reCHInnecting

Google

UCLA HCI RESEARCH

AR

I was doing a self-introduction in a group social event, and it really attracted people's attention and increased the fun and engagement at the beginning

P17

“

It helped understand unfamiliar words in English as a non-native speaker. E.g., Andromeda

”

I use automatic (auto-display) mode all the time, but change to on-demand mode in important meetings because I don’t want to interrupt other speakers

“

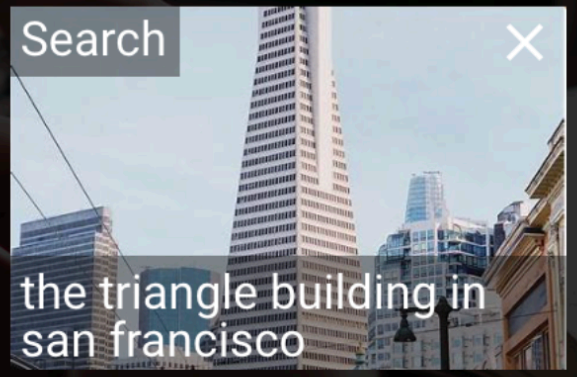
VC pops up images for words that I don’t understand, like ‘groomhaven’, ‘borg sphere’ in a social meetup

”

5. Ambiguous Reference

“My favorite snack is some kind of blue potato chips” →

“You know, the triangle building in San Francisco” →



A. Transcription Parsing

"I just went to Disneyland with my family last weekend, it was super fun! And..."

Speech-to-Text

100 ms interval

"I just went to Disneyland with my family last weekend, it was super fun!"

Retrieve last complete sentence and last sentence > n_min

Fine-tuned Language Model

B. Visual Intents Prediction

<A photo> *of* ***<Disneyland>*** *from* ***<online image search>***

<A photo> *of* ***<me and my family at Disneyland last weekend>*** *from* ***<personal album>***

<A emoji> *of* ***<happy face>*** *from* ***<emoji search>***

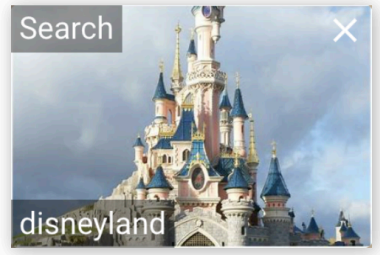
Predict Visual Intents

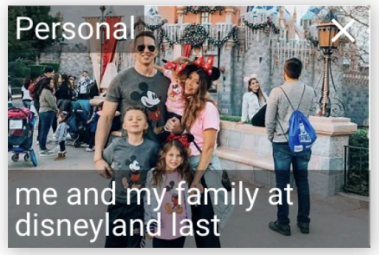
C. Visual Generation

Online Image Search

Text Image Retrieval

Custom Database





Retrieve Relevant Visuals

System – Model

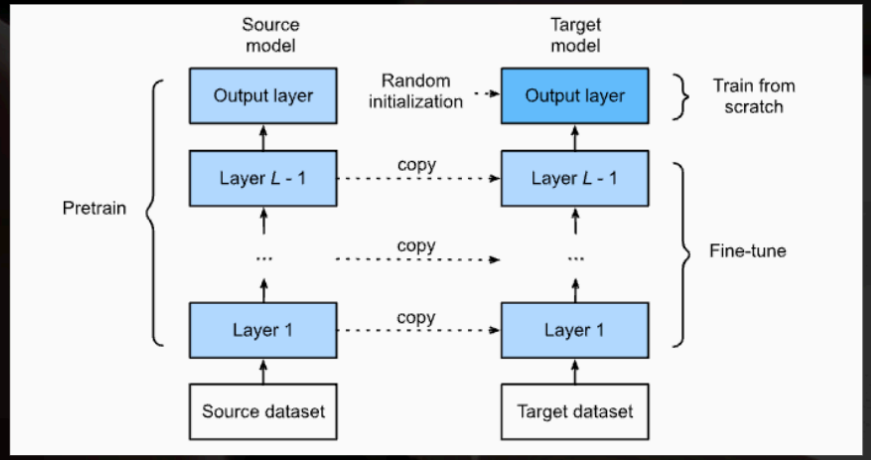
1. Processed crowd workers' responses into

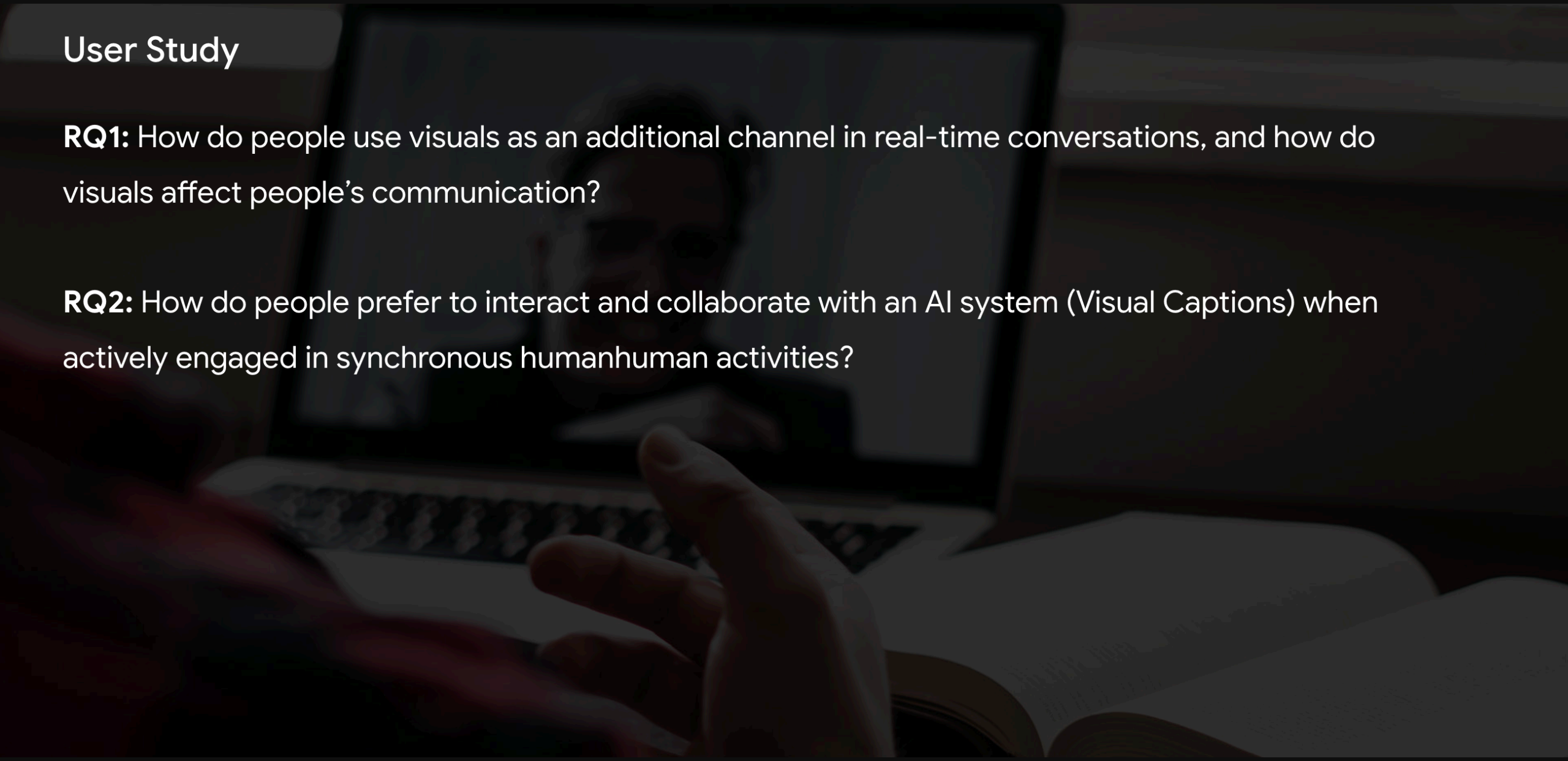
   `<visual type> of <visual content> from <visual source>`

2. Parsed as input data for pre-trained language models (GPT3, LaMDA)

   ```
   {"prompt":"Thanks great presentation! ->",
   "completion":" emoji of hand clapping from emoji search\n"}
   ```

3. Fine-tuning / transfer learning





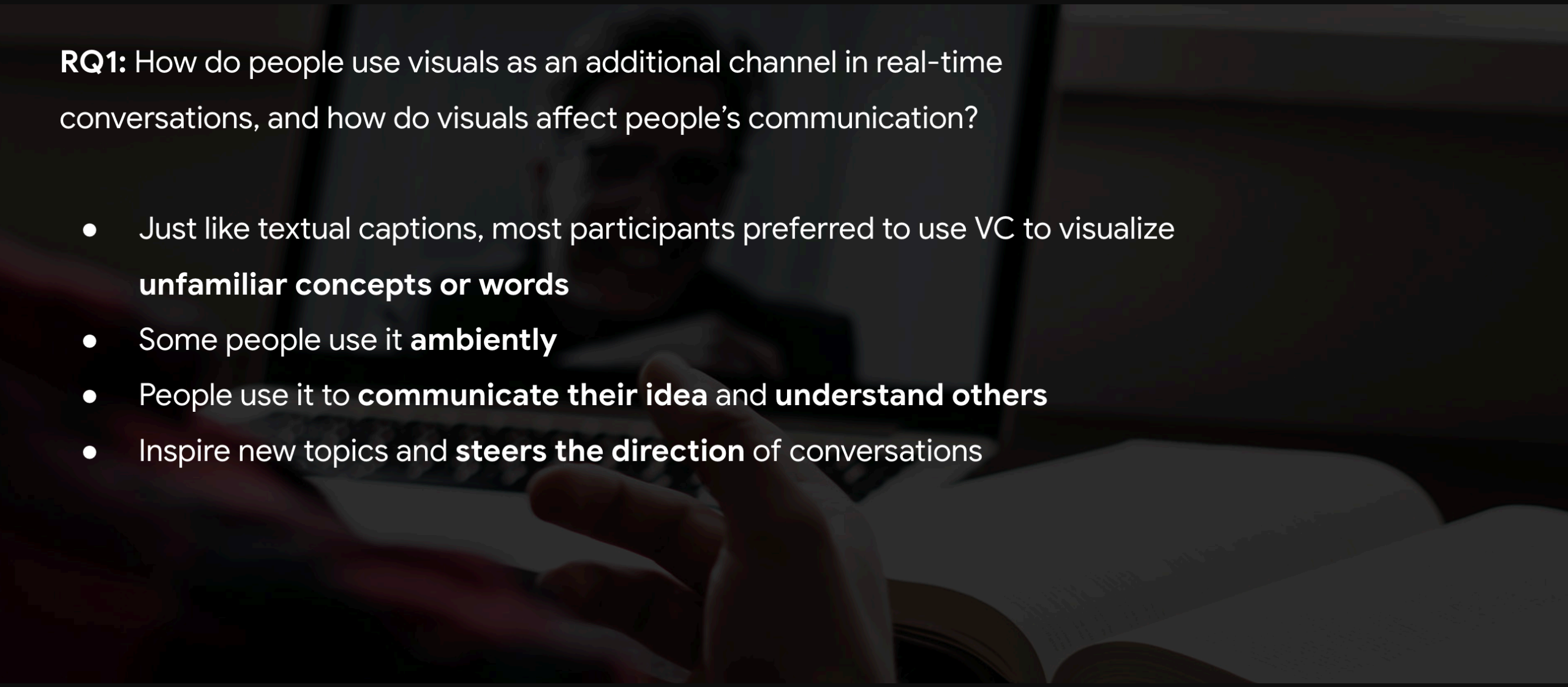
User Study

RQ1: How do people use visuals as an additional channel in real-time conversations, and how do visuals affect people's communication?

RQ2: How do people prefer to interact and collaborate with an AI system (Visual Captions) when actively engaged in synchronous humanhuman activities?

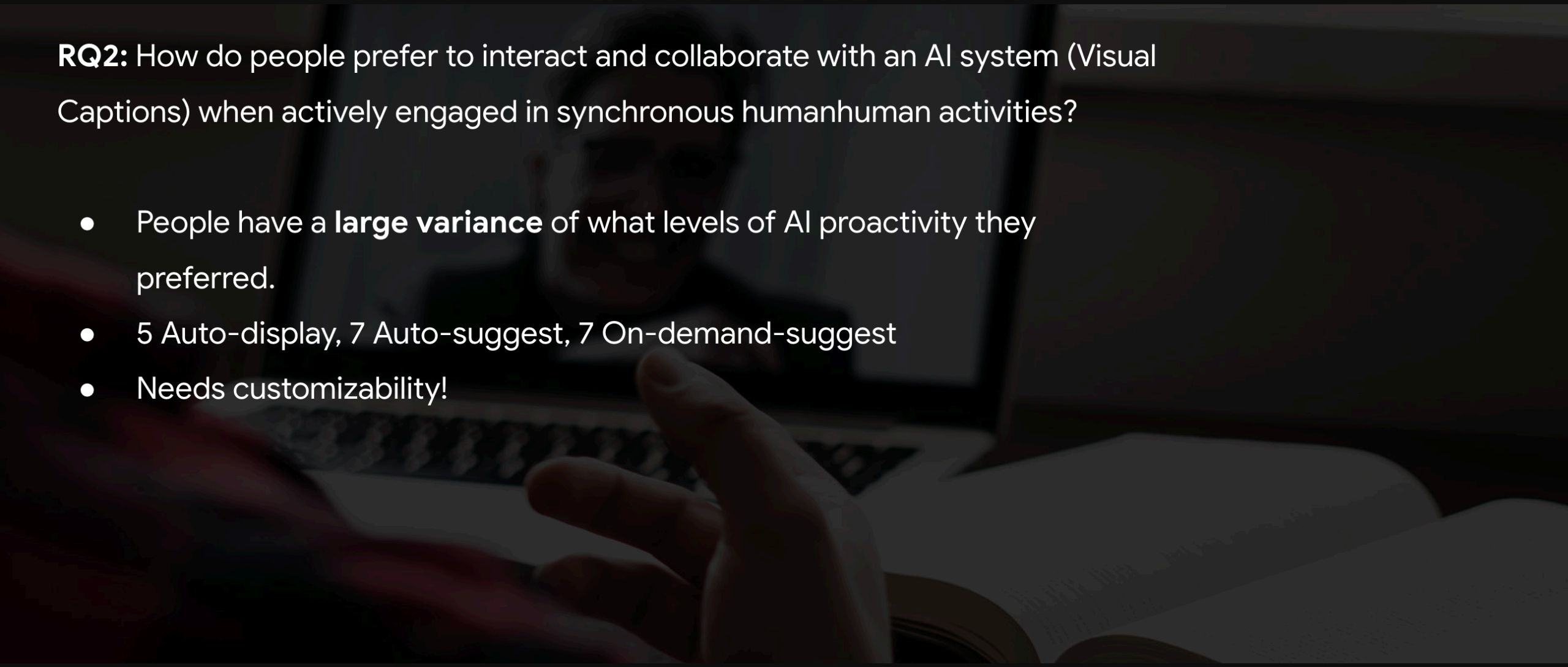
User Study

- 20 participants (9 female and 11 male)
- 21 – 61 years old
- 4 scripted conversations + 5-10 minutes open-ended conversation
- 3 levels of AI proactivity
  - **Auto-display**
  - **Auto-suggest**
  - **On-demand-suggest**
- TLX index, Likert scale ratings, semi-structured interviews
- Deployment study with 10 participants WIP



RQ1: How do people use visuals as an additional channel in real-time conversations, and how do visuals affect people's communication?

- Just like textual captions, most participants preferred to use VC to visualize **unfamiliar concepts or words**
- Some people use it **ambiently**
- People use it to **communicate their idea** and **understand others**
- Inspire new topics and **steers the direction** of conversations

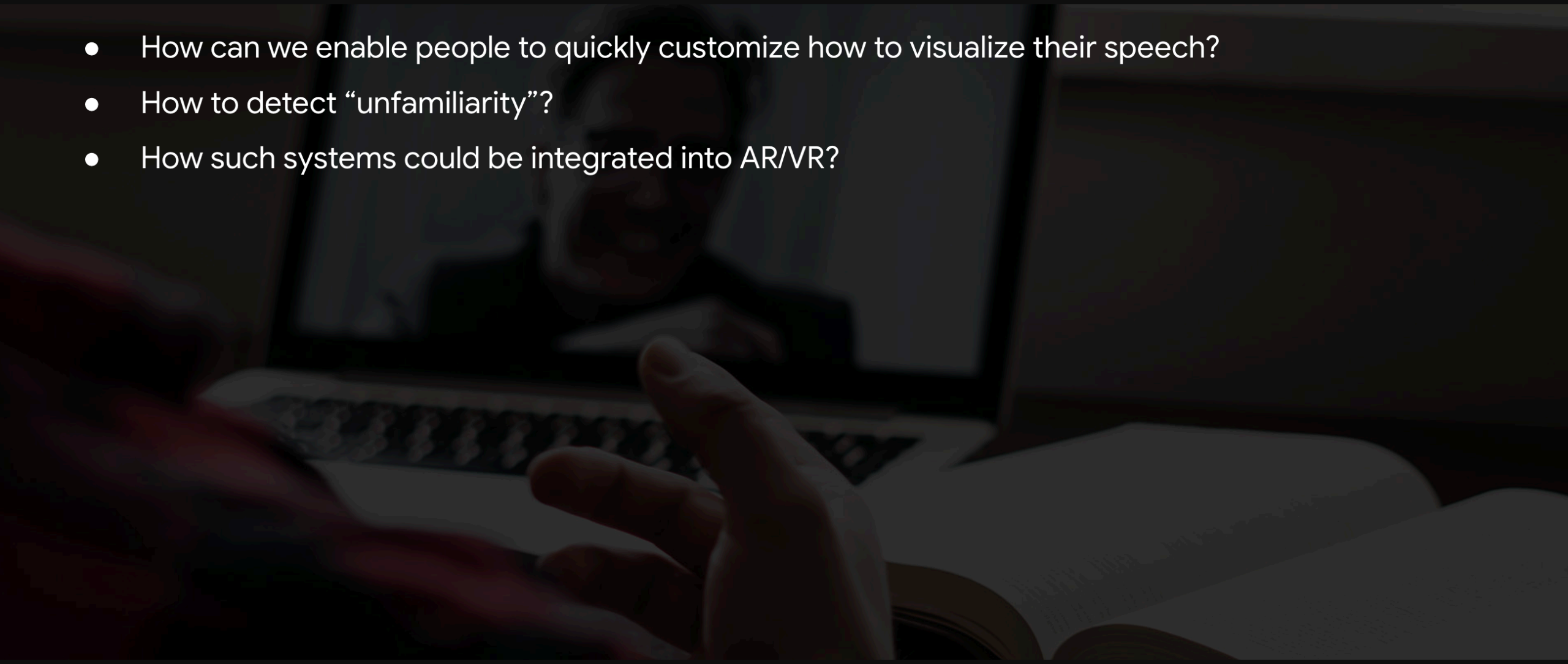


RQ2: How do people prefer to interact and collaborate with an AI system (Visual Captions) when actively engaged in synchronous humanhuman activities?

- People have a **large variance** of what levels of AI proactivity they preferred.
- 5 Auto-display, 7 Auto-suggest, 7 On-demand-suggest
- Needs customizability!

RQ2: How do people prefer to interact and collaborate with an AI system (Visual Captions) when actively engaged in synchronous humanhuman activities?

- **Auto-display**
  - Pros: Minimum interaction needed
  - Cons: Less human control & potential infringement of privacy
- **Auto-suggest**
  - Pros: Less interaction, understand system's capabilities
  - Cons: Continuous suggestions may be distracting
- **On-demand-suggest**
  - Pros: Least distracting
  - Cons: More interaction required



- How can we enable people to quickly customize how to visualize their speech?
- How to detect “unfamiliarity”?
- How such systems could be integrated into AR/VR?

I feel like we were responding to the photos. When we were talking about a whale watching tour, it suggested an image of people on a very small boat. We got to further discuss what boat I was on in the tour and so on.

“When I would really want visuals is like people don’t know what I was talking about. For example when I just mentioned Santa Monica Pier, it’s great that I can easily explain what it is.”

“The system is especially helpful when it shows me something I don’t know, like in this example it shows me a picture of Rodeo Drive. Whenever I’m confused I can just take a look at the right side and see intuitively what they are”

P10

Back in the beginning when we were talking about the Avatar, there are like four or five different versions we might be discussing, the picture helped crystallize it instantly

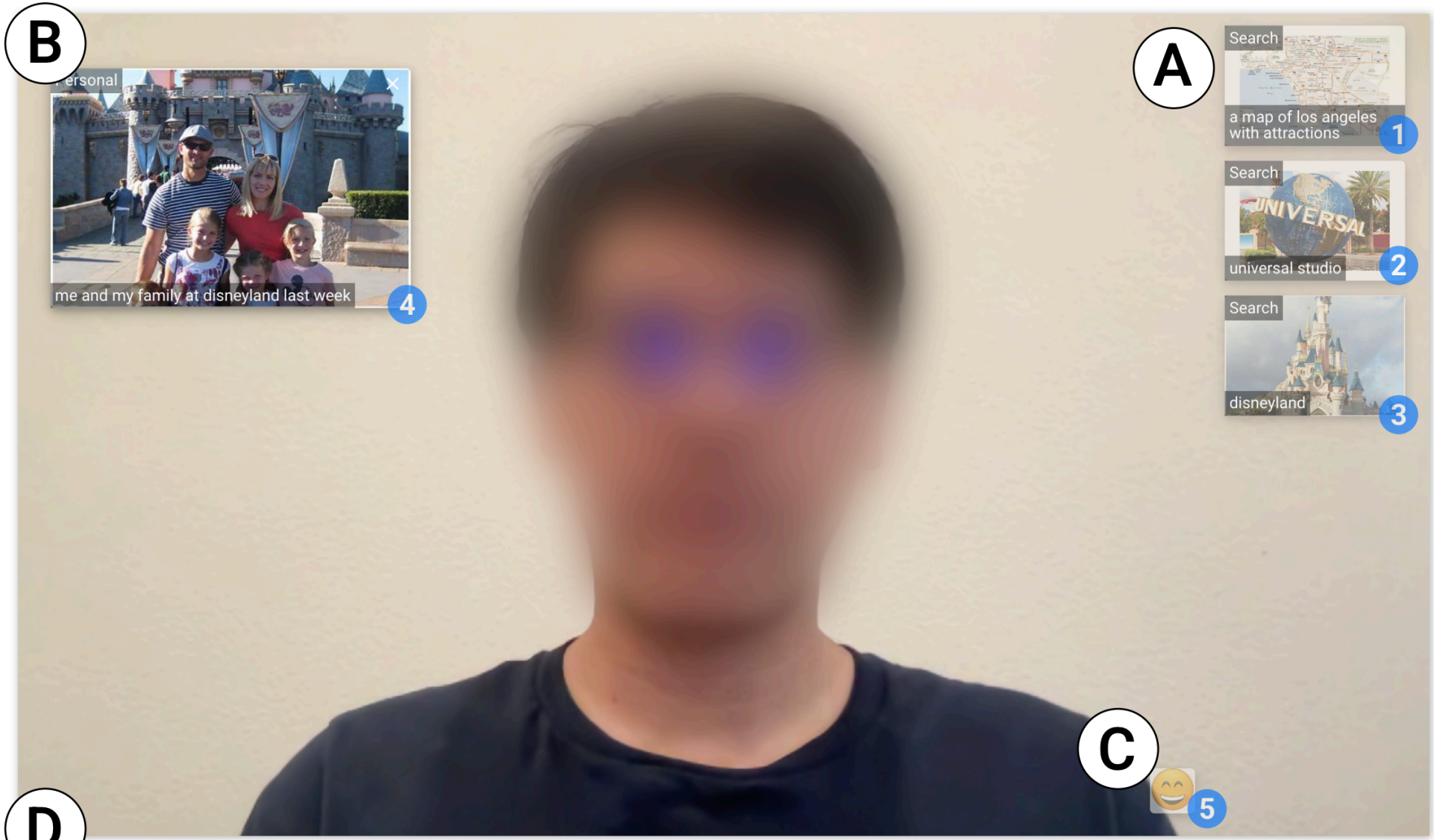
When chatting with recommended pictures, our interaction has increased. The conversation is getting longer with more content

“

It takes some time to identify which images are proper to share. It interrupts the conversation a little bit, but otherwise I feel OK

”

P7



If you're visiting LA, You should definitely visit Universal Studio and Disneyland.

1 2 3

I just went to Disneyland with my family last weekend, and it was so much fun!

4 5

Instruction

- Please carefully read the instruction and examples. Your HIT will be **rejected** if you do not follow the intstructions.
- Detemine what **visual content (e.g. images, photos, 3D objects, gifs, videos, visual effects)** could be shown to supplement the last sentence, given context and previous conversation
- Answer in the format of *Visual types* of *information to visualize*
- Seperate your answer by ";" if there are **multiple** visuals that could be added.
- "Context" and "Previous" are just provided as contextual information, please only add visuals to supplement the last sentence.
- Type "**none**" if you think no visual is appropriate for this sentence.

Examples

Context: talking about where to visit in LA.
Previous: "*So, what do you want to do while you're here? Well, there's plenty to see.*"
Last Sentence: "*If you're interested in Hollywood, you could visit the Walk of Fame, Rodeo Drive, Grauman's Chinese Theatre.*"
Visuals to supplement the last sentence: A photo of Hollywood; A photo of Walk of Frame; A photo of Rodeo Drive; A photo of Chinese theatre.

Context: chatting about what did people do last weekend
Previous: "*What did you do last weekend? Sounds like you had lots of fun.*"
Last Sentence: "*I went to Disneyland with my family last weekend.*"
Visuals to supplement the last sentence: A photo of me and my family in Disneyland last week.

Context: chatting when having dinner
Previous: "*How's the chicken?*
Last Sentence: *It's delicious!.*"

Visual Captions Settings

Enable Visual Captions ☑

AI Proactiveness: Suggestion

Advanced Settings

Algorithm

All Participants' Captions ☐

Suggest Emojis ☑

Suggest Personal ☑

Model: Most Capable

Min num words: 20

Last N Sentences: 1

Scrolling View

Max Visuals: 5

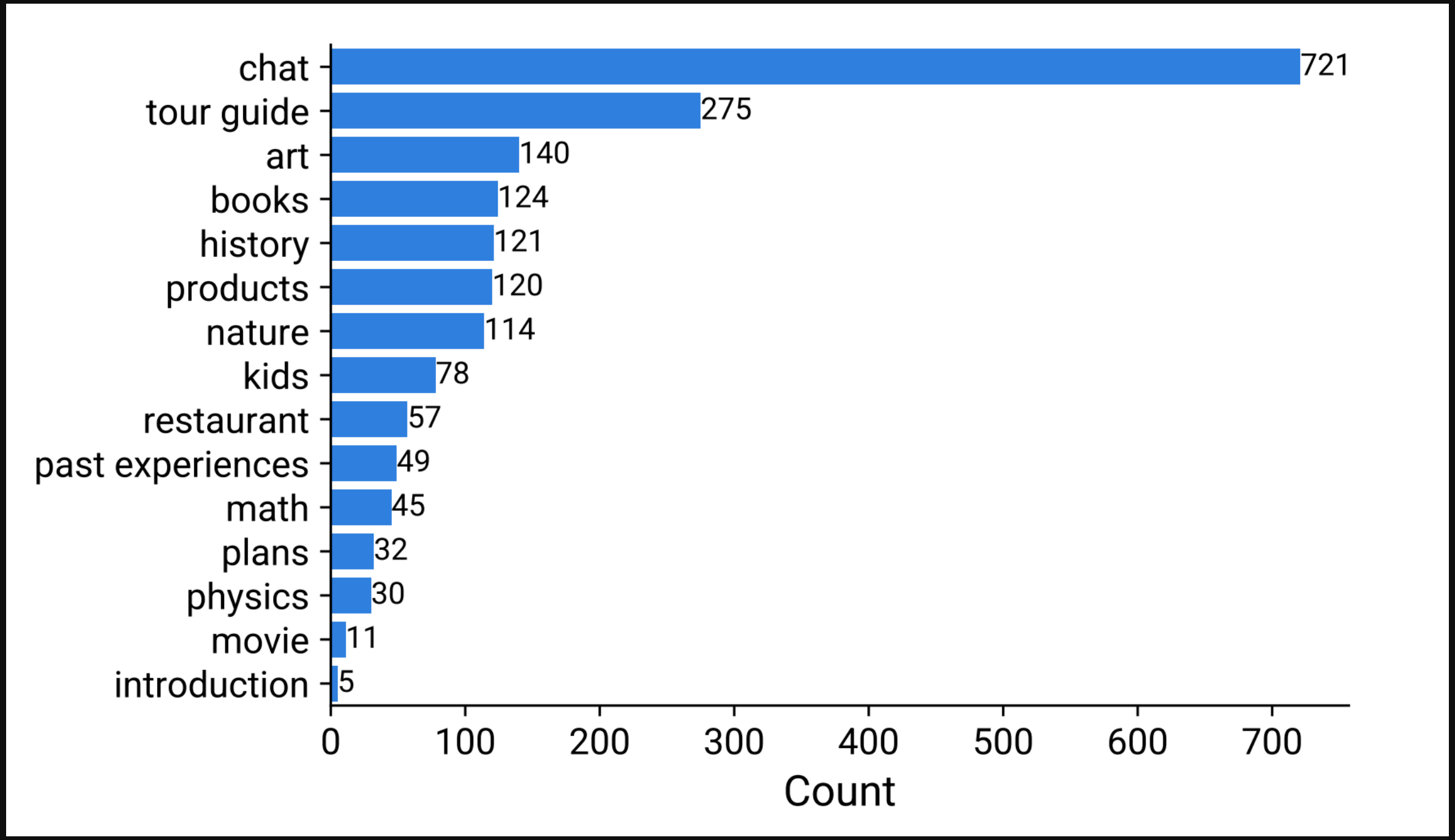
Max Emojis: 4

Visual Size: 1

Logging

Enable Logging ☐

Download Log


chat
tour guide
art
books
history
products
nature
kids
restaurant
past experiences
math
plans
physics
movie
introduction
721
275
140
124
121
120
114
78
57
49
45
32
30
11
5
0
100
200
300
400
500
600
700
Count

Visual Captions

Augmenting Verbal Communication with On-the-fly Visuals

Xingyu “Bruce” Liu, Vladimir Kirilyuk, Xiuxiu Yuan,
Alex Olwal, Peggy Chi, Xiang "Anthony" Chen, Ruofei Du

github.com/google/archat

CHI23
Hamburg, Germany | Hybrid
April 23-28, 2023
reCHInnecting

Google

UCLA HCI RESEARCH

AR

Visual Captions

Augmenting Verbal Communication with On-the-fly Visuals

Xingyu “Bruce” Liu, Vladimir Kirilyuk, Xiuxiu Yuan,
Alex Olwal, Peggy Chi, Xiang "Anthony" Chen, Ruofei Du

github.com/google/archat

CHI23
Hamburg, Germany | Hybrid
April 23-28, 2023
reCHInnecting

Google

UCLA HCI RESEARCH

Opensourced at https://github.com/google/archat

Visual Captions: Augmenting Verbal Communication with On-the-fly Visuals

Xingyu “Bruce” Liu, Vladimir Kirilyuk, Xiuxiu Yuan, Alex Olwal, Peggy Chi, Xiang "Anthony" Chen, Ruofei Du

UCLA Google

UCLA HCI RESEARCH