Experiencing Visual Captions: Augmented Communication with Real-time Visuals using Large Language Models

Xingyu 'Bruce' Liu
UCLA
Los Angeles, CA, USA
xingyuliu@ucla.edu

Vladimir Kirilyuk
Google Research
Mountain View, CA, USA
vkyryliuk@google.com

Xiuxiu Yuan
Google Research
Mountain View, CA, USA
xiuxiuyuan@google.com

Peggy Chi
Google Research
Mountain View, CA, USA
peggychi@google.com

Xiang 'Anthony' Chen
UCLA
Los Angeles, CA, USA
xac@ucla.edu

Alex Olwal
Google Research
Mountain View, CA, USA
olwal@acm.org

Ruofei Du*
Google Research
San Francisco, CA, USA
me@duruofei.com

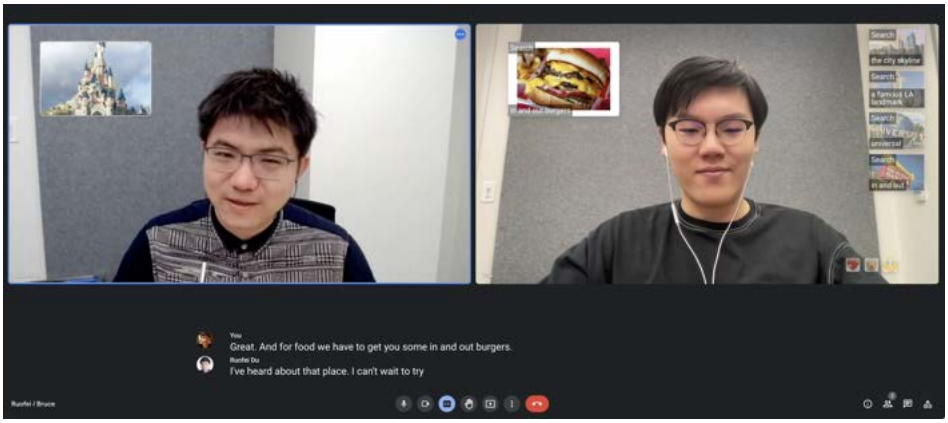

Figure 1: Visual Captions is a real-time system that suggests relevant visuals in conversations. Our interface built as a Chrome extension allows users to share visuals on-the-fly in video conferencing platforms such as Google Meet.

ABSTRACT

We demonstrate Visual Captions, a real-time system that integrates with a video conferencing platform to enrich verbal communication. Visual Captions leverages a fine-tuned large language model to proactively suggest visuals that are relevant to the context of the ongoing conversation. We implemented Visual Captions as a user-customizable Chrome plugin with three levels of AI proactivity: *Auto-display* (AI autonomously adds visuals), *Auto-suggest* (AI proactively recommends visuals), and *On-demand-suggest* (AI suggests visuals when prompted). We showcase the usage of Visual Captions in open-vocabulary settings, and how the addition of visuals based on the context of conversations could improve comprehension of complex or unfamiliar concepts. In addition, we demonstrate three approaches people can interact with the system with different levels of AI proactivity. Visual Captions is open-sourced at https://github.com/google/archat.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; *Mixed / augmented reality*.

KEYWORDS

augmented communication, large language models, video-mediated communication, online meeting, collaborative work, dataset, text-to-visual, AI agent, augmented reality

ACM Reference Format:
Xingyu 'Bruce' Liu, Vladimir Kirilyuk, Xiuxiu Yuan, Peggy Chi, Xiang 'Anthony' Chen, Alex Olwal, and Ruofei Du. 2023. Experiencing Visual Captions: Augmented Communication with Real-time Visuals using Large Language Models. In *The 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23 Adjunct), October 29-November 1, 2023, San Francisco, CA, USA.* ACM, New York, NY, USA

1 INTRODUCTION

Recent advances in video conferencing have significantly improved remote video communication through features like live captioning

*Corresponding author. Work was done while Xingyu 'Bruce' Liu was a Student Researcher at Google. Contact: xingyuliu@ucla.edu and me@duruofei.com.

UIST '23 Adjunct, October 29-November 1, 2023, San Francisco, CA, USA

ACM ISBN 979-8-4007-0096-5/23/10.
https://doi.org/10.1145/3586182.3615978

and noise cancellation. However, there are various situations where dynamic visual augmentation would be useful to better convey complex and nuanced information. For example, when talking about a recent trip, people may use photos from their album to help their listeners follow along.

In this demonstration paper, we present Visual Captions [1], a real-time system designed to address the challenges of augmenting synchronous human-human verbal communication with visuals. Visual Captions automatically predicts the "visual intent" of a conversation, or the visuals that people would like to show at the moment of their conversation, and suggests them for users to immediately select and display. We trained an accurate, robust, and open-vocabulary language model to predict "visual intents" in conversations, achieving an 86.59% validation token accuracy. Based on the language model, we implemented Visual Captions as a user-customizable Chrome plugin with three levels of AI proactivity: *Auto-display* (AI autonomously adds visuals), *Auto-suggest* (AI proactively recommends visuals), and *On-demand-suggest* (AI suggests visuals when prompted). We envision Visual Captions to facilitate live conversations in multiple ways, including helping to explain and understand unfamiliar concepts, clarify language ambiguities, and make conversations more engaging.

2 SYSTEM OVERVIEW

We developed Visual Captions on the base of *ARChat*, a web-based rapid prototyping platform we built where developers can quickly build and deploy augmented communication systems. We additionally built a *settings page* for users to customize the visual types to generate, suggestion modes of the AI, and the visual layout.

2.1 Visual Captions System

2.1.1 Generating Visual Suggestions. Visual Captions automatically suggests relevant visuals based on users' conversation content (Figure 3). Our system continuously retrieves the automatic captions from Google Meet, and queries for a window of captions every 100 ms. The queried caption is pre-processed and sent as the input to the visual intent prediction model. This query window is customizable on the settings page (A.2). By default, our system queries the last two sentences, signified by end-of-sentence punctuation (".", "?", or "!"). To enable responsive visuals for incomplete sentences (*e.g.*, *"Andy Warhol is one of"*), our system also queries visuals if it has more than $n_{\min}$ words ($n_{\min} = 4$ by default).

The model predicts (1) the information to visualize, (2) the type of visual to present, and (3) the source for the visual. For example, it may suggest visualizing *"Santa Monica pier at night"* with an *image* from *online search*. The returned information initiates different pipelines based on the predicted visual type and source. For example, if the model prediction returns *"A photo of me and my family at Disneyland from personal album"*, our system will run a personal album search and return the photo with the highest CLIP score, *i.e.* a strong relationship between the language in the text and the visual information in the image. If the model predicts *"a map of Los Angeles from online search"*, our system will run an online image search with the search term *"A map of Los Angeles"*.

Visual Captions then creates a *Visual Widget* object which contains attributes *imgURL* (the URL of the retrieved visual), *description* (the search term), *visual source*, and *visual type*. Widgets are rendered as an HTML element with the visual, the description in the bottom-left corner and the source in the top-left corner (Figure 2), and added to the video conference interface. Visual widgets are by default 50% transparent (customizable setting) to make them more ambient and less distracting to the main conversation, and change to non-transparent on hover.

2.1.2 Scrolling View. In *auto-suggest* and *on-demand-suggest* modes where users have to explicitly approve our systems' visual suggestions, all generated visual widgets are first displayed in a *Scrolling View* (Figure 2A). Visual suggestions in the scrolling view are private to the users and not shown to others in the meeting. The scrolling view is automatically updated when new visuals are suggested by the system, and removes the oldest visual widget if it exceeds the maximum amount (customizable *# Max Visuals* in subsection A.2). Similarly, emojis are displayed in a separate scrolling view in the bottom right corner of the screen (Figure 2C). For example (Figure 2), when a user said *"If you're visiting LA, you should definitely visit the Disneyland and Universal Studios. I just went to Disneyland with my family last weekend, it was super fun!"*, Visual Captions suggests and adds to the scrolling view, in order:

(1) "A map of Los Angeles with highlighted attractions *from online image search*"
(2) "Disneyland *from online image search*"
(3) "Universal studios *from online image search*"
(4) "A photo of me and my family at Disneyland last weekend *from my photo album*"
(5) "A happy face emoji"

2.1.3 Spotlight View. To make an image visible to all parties, the user may click the widget to move it to the *Spotlight View* (Figure 2B). Visuals in the spotlight view can be moved, resized, and deleted.

2.2 AI Proactivity in Visual Captions

Our system provides three levels of AI proactivity:

Auto-display (high-proactivity). In the auto-display mode, the system autonomously searches and displays visuals publicly to all meeting participants. AI has full control and no interaction is needed. The scrolling view is disabled.

Auto-suggest (medium-proactivity). In the auto-suggest mode, the suggested visuals will be shown in the private scrolling view. A user then click's a visual to display it publicly. In this mode, the AI is proactively recommending visuals, but the user selects when and what to display.

On-demand-suggest (low-proactivity). In the on-demand-suggest mode, the AI will only suggest visuals if a user taps the space bar. The system immediately queries the captions and stays on for 3 seconds to query the following speech.

2.3 Use Cases and Scenarios

We share a sample of use cases and scenarios where Visual Captions could be helpful to supplement conversations.

Education and Lectures. Presentation plays a vital role in education, but may not cover everything the lecturer is talking about. Oftentimes, when a student asks an out-of-scope question or the

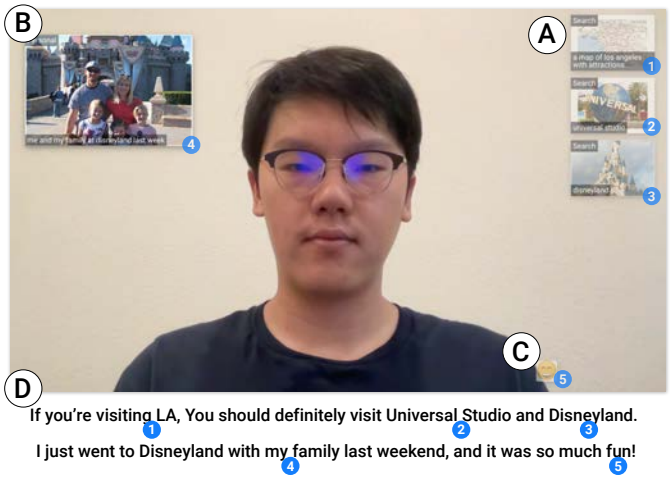

Figure 2: In Visual Captions's interface (default auto-suggest mode), the *scrolling view* (A) displays privately candidates of visual suggestions generated by our visual intent prediction model. Emoji suggestions are displayed on the bottom right corner (C). Users can click and display the visual to the *spotlight view* (B) to share it publicly.

teacher talks about a new concept, a visual augmentation system may help visualize the key concepts or unfamiliar words in the conversation.

Casual Conversations. We envision that real-time visual augmentation may enhance casual chats by providing more information to the conversation, for example, bringing up personal photos, visualizing unknown dishes, and instantly fetching movie posters.

Business and Utility. In business meetings, a private visual channel can remind people of unfamiliar faces when their names are called out. As a utility tool, a screenshot of Google Maps can guide users' navigation; a call out of items may remind users where they are in a memory palace.

Creativity. As a creativity tool, a speech-driven generative text-to-image model can help people with brainstorming, create initial design draft, or efficiently generate mind maps.

Storytelling. Speech-to-image tools also have potential to tell stories. For example, when speaking about animal characters, show life-size 3D animals in augmented reality displays.

3 CONCLUSION

In this demonstration paper, we introduced Visual Captions, a system designed to augment synchronous human-human verbal communication with visuals, by predicting the "visual intent" of a conversation and suggesting relevant visuals for users to immediately display. We believe that incorporating visual augmentations in conversations could greatly benefit communication, particularly in the context of conveying complex, nuanced, and unfamiliar information. We envision Visual Captions to be a valuable tool for improving understanding and engagement in live conversations.

Word Count: 1464 words.

REFERENCES

[1] Xingyu "Bruce" Liu, Vladimir Kirilyuk, Xiuxiu Yuan, Alex Olwal, Peggy Chi, Xiang "Anthony" Chen, and Ruofei Du. 2023. Visual Captions: Augmenting Verbal Communication with On-the-Fly Visuals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 108, 20 pages. https://doi.org/10.1145/3544548.3581566

A SYSTEM DETAILS

A.1 ARChat: A Rapid Prototyping Platform for Augmented Communication

We developed ARChat to enable large-scale and long-term deployment of augmented communication prototypes. ARChat is a rapid prototyping framework written in TypeScript and JavaScript with native support for speech-to-text, TensorFlow.js, 3D rendering. It can also be deployed as a Chrome browser plugin. In contrast to former systems that leverage WebRTC protocols for custom augmented communication prototypes, ARChat integrates with existing videoconferencing platforms (*e.g.*, Google Meet, Zoom, Microsoft Teams) by simulating a virtual camera to process audiovisual sources and render augmented views. We fetch the video stream of the user's selected camera, render the frame to an off-screen canvas with our augmented content, then stream the canvas to the simulated virtual camera via a local WebRTC stream. When used as a Chrome plugin, ARChat also supports fetching cloud-based subtitles from videoconferencing platforms (*e.g.*, Google Meet) to leverage state-of-the-art web-based speech-to-text. For this work, ARChat has facilitated the development and deployment of Visual Captions.

A.2 Visual Captions Settings

The Visual Captions settings page (Figure 4) allow users to fully customize how they prefer to control and display AI-suggested visuals. For system functionality, users can enable or disable Visual Captions, Emojis, and visuals from personal albums. Users can additionally control what prediction model to use (from "Most capable, but slower" to "Fastest, but less capable"), and how the system queries for visuals (after certain number of words, or after

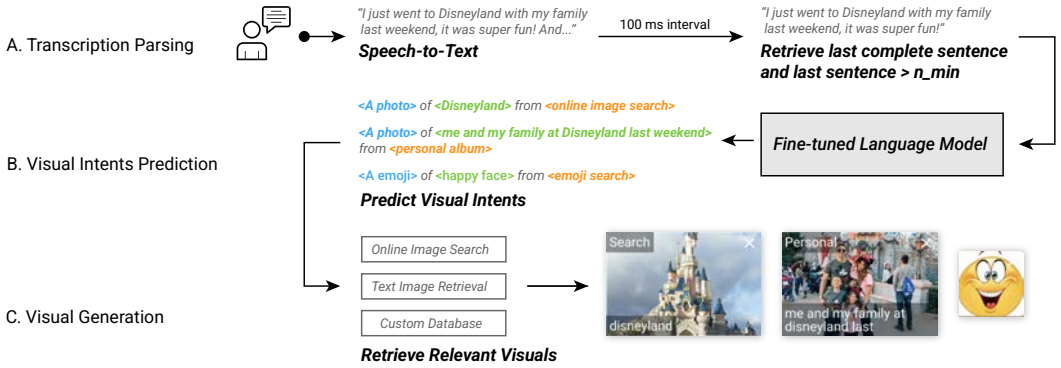

Figure 3: System workflow of Visual Captions. The workflow consists of three major steps: A. Transcription Parsing; B. Visual Intents Prediction; C. Visual Generation.

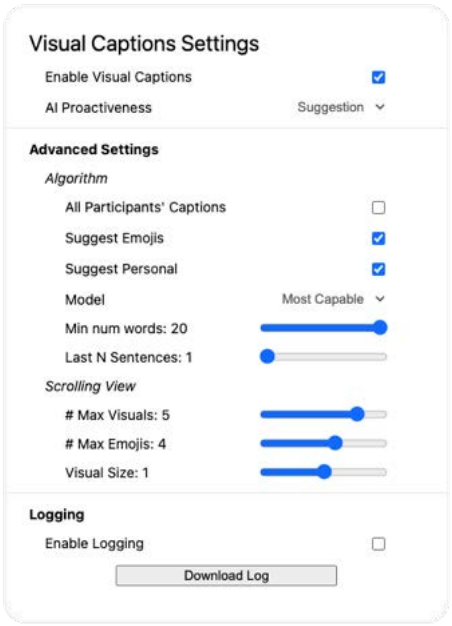

Figure 4: Visual Captions allows users to customize settings including levels of AI proactivity, whether to suggest emoji or personal images, punctuality of visual suggestions, visual suggestion models, etc.

a complete sentence). The system can also suggest visuals based on other meeting participants' speech if enabled (*All Participants' captions*). The layout of Visual Captions on Google Meet is also customizable on the settings page.

A.3 System Diagram

Figure 3 shows the computational pipeline of the Visual Captions system.