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

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

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

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

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

Ruofei Du
Google Research
San Francisco, CA, USA
ruofei@google.com

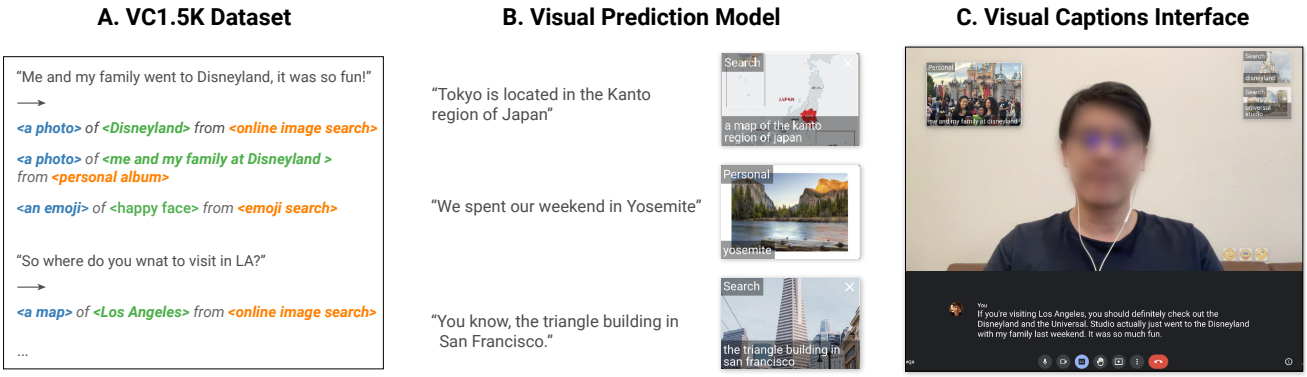

Figure 1: Visual Captions is a real-time system that suggests relevant visuals in conversations. We contribute (A) VC1.5K, a crowdsourced dataset that contains 1595 quadruples of language, visual content, type, and source; (B) a visual prediction model fine-tuned on GPT-3 to suggest relevant visuals, and (C) Visual Captions interface that allows users to share visuals on-the-fly in video conferences.

ABSTRACT

Video conferencing solutions like Zoom, Google Meet, and Microsoft Teams are becoming increasingly popular for facilitating conversations, and recent advancements such as live captioning help people better understand each other. We believe that the addition of visuals based on the context of conversations could further improve comprehension of complex or unfamiliar concepts. To explore the potential of such capabilities, we conducted a formative study through remote interviews (N=10) and crowdsourced a dataset of over 1500 sentence-visual pairs across a wide range of contexts. These insights informed 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 relevant visuals in open-vocabulary conversations. We present findings from a lab study (N=26) and an in-the-wild case study (N=10), demonstrating how Visual Captions can help improve communication through visual augmentation in various scenarios.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**.

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, 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 (CHI '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3581566

1 INTRODUCTION

Recent computer-mediated systems are increasingly facilitating verbal communication, which is often the predominant mode of communication [3, 49]. Platforms like Google Meet, Zoom, and Microsoft Teams, have been widely adopted and provided capabilities such as live captioning and noise cancellation to facilitate conversations.

While such tools help people understand each other better, we envision that visual augmentations that leverage the semantics of spoken language could also be helpful, especially for conveying complex, nuanced, and unfamiliar information. People are already using visual aids to provide additional context and clarification in daily conversations. For example, when talking about a recent trip, people may use photos from their album to help their listeners follow along. Similarly, when describing a new restaurant to a friend, one might say "*it's a small place with a lot of character*" and then search for a picture online to show what the place looks like. Research has shown that people learn more effectively from videos than from audio-only versions of the same material [22, 29, 47], and prefer podcasts [51] and stories [25] with visuals over those without. The Multi-modal Phenomena [30, 44, 46] and the Principle of Inverse Effectiveness [45] have also proved that the human sensory system has a superadditive effect when responding to stimulus from multiple, simultaneous modalities. As the adage goes, *"a picture is worth a thousand words"*.

Prior research has proposed various automated text-to-visual systems, that can automatically transform audio-only content into audiovisual content [25, 51], and create realistic images and art from natural language [32, 54]. However, augmenting *synchronous human-human verbal communication* with visuals presents unique challenges that have not been addressed by these systems.

First, the input is a continuous stream of conversation, rather than a discrete textual description of the visual, requiring the system to go beyond keyword-based approaches, like named entity detection, to understand the *implicit* intent of what people want to show in the context. *Second*, when people are actively engaged in conversation, they have limited cognitive resources to interact with AI prompts and results, making it necessary for the interaction to be subtle and minimal to avoid disrupting the conversation. *Third*, without a real-time system deployed, it is difficult to study how people could interact with and benefit from visuals in real conversations, and how such systems would impact their communication.

In this work, we introduce Visual Captions, a 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 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 conducted two rounds of formative studies with 10 participants and collected, annotated, and analyzed over 1,500 dialogues from 246 crowdworkers to understand the interest in such capabilities. Using this dataset, 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 evaluated Visual Captions with 20 participants in a controlled lab study where participants chatted casually in four different scenarios, Participants found real-time visuals facilitated live conversations in multiple ways, including helping to explain and understand unfamiliar concepts, clarify language ambiguities, and make conversations more engaging. Participants also reported diverging preferences in how to interact with the AI in-situ. We also conducted a two-week in-the-wild deployment study with 10 participants to allow people to use the system in their everyday conversations. Participants reported that different levels of AI proactivity in Visual Captions were preferred in various social scenarios.

In summary, we contribute:

- A dataset of 1595 visual intents collected from 246 crowd workers, covering 15 categories of topics.
- A design space for real-time visual augmentation of verbal communication, derived from findings from two brainstorming sessions in a formative study (N=10).
- A visual intent model that predicts the type, source, and content for the visuals that people may want to display in conversations.
- Visual Captions, an open-source Chrome plugin for real-time augmented visuals in video conferencing products[1].
- Results from a user study (N=26) and a deployment study (N=10) that demonstrates the potential benefits of using visuals to augment conversations.

2 RELATED WORK

Our work is inspired by prior research in automatic text-to-visual enhancement, proactive AI agents, and augmented communication.

2.1 Text-to-Visual Systems

Researchers have developed various computational techniques to use visual materials to support text-based contents [24, 36, 37, 40, 52, 56]. Early work enhanced emails by suggesting relevant personal photos related to the email topics [23]. As text chats became more common, real-time suggestions were developed to assist users in adding illustrations [21], images [10], and animated graphics [27, 50, 55]. For a comprehensive review of text-to-visual systems, we refer readers to the surveys by Hassani et al. [15] and Zakraoui et al. [53].

Recent research in HCI converts text articles to audio-visual slideshows [7, 25] and automatically generates visuals for travel podcasts [51] that consider sequential presentation and storytelling. When the source documents contain multimedia materials, automatic methods could crop, zoom, or loop images and videos to enhance text content [8, 9]. These methods demonstrated novel approaches for enriching text-driven experiences. Our work builds on these systems but for face-to-face conversations, and providing real-time visuals relevant to the spoken content.

[1]Visual Captions is open-sourced at: https://liubruce.me/visual_captions

2.2 Proactive and Continuous AI Agent

Prior literature has investigated both proactive and continuous AI agents. In 1996, Rhodes et al. [39] propose one of the earliest system for continuously providing relevant information by monitoring human activities. However, AI is never perfect in real-world deployment [23]. Meurisch et al. [31] present an in-the-wild study of proactivity levels in voice-controlled assistants. Andolina et al. [2] present SearchBot, which provides continuous recommendations of related documents and entities in a non-intrusive way [1] during voice conversations. Their work leveraged Google's Cloud Natural Language API for extracting recognized entities from the transcripts. In contrast to their work, our system fine-tuned a large language models and better understand conversations with visual intent. We elaborate the differences between our approach and key-word based approach in Section 5.

In addition, mixed-initiative interactions, where both the user and the computer can initiate actions and take turns controlling the flow of the interaction, have been a topic of interest in HCI research. Horvitz et al. [19] reviewed critical factors for the effective integration of human and AI. Our work is inspired by prior research in mixed-initiative interaction and proactive AI. We provide three levels of AI proactivity and investigate the effects of these levels on users' perception and acceptance of the AI agent in a real-world setting. In addition, we study people's preferences on mixed-initiative interactions when they are actively engaged in conversations that limit their cognitive load and interaction bandwidth.

2.3 Augmented Communication

Our work is related to systems that augment verbal communication with text, interactive graphics, and computer-assisted actions. With recent advances of real-time machine learning techniques and wearable displays, there is a grouping trend of integrating augmented communication in everyday conversations [17]. Lyons et al. [28] propose Dual–purpose Speech, which automatically use speech to navigate a user's calendar, save transient information, and create asynchronous tasks. Saquib et al. [42] enable interactive presentations with body-driven graphics, by mapping a variety of body movements to a wide range of graphical manipulations beforehand. Müller et al. [33] present Cloudbits, which interactively renders calendar, emails, hotel, and restaurant during conversations via a Wizard-of-Oz study on an HMD. Peng et al. [35] present SpeechBubbles, a real-time speech-to-text system that associates utterances with speakers. Their studies show that participants significantly prefer their layout over traditional captions. Ogawa et al. [34] propose a smartwatch-based approach to suggesting topics in casual conversations. Social messaging apps like Snapchat and Facebook Messenger also suggest pre-defined stickers and emojis through keyword detection.

Visual Captions instead uses a fine-tuned large language model to proactively suggest relevant visuals in open-vocabulary conversations. Our system provide personalized visuals in real-time that are tailored to the specific context. To the best of our knowledge, Visual Captions is the first real-time system trying to understand user's visual intent, suggest relevant visual content, source, type, offers different levels of proactivity, and gets deployed in the wild without the limitation of pre-scripted conversations.

3 FORMATIVE STUDY

To understand how people would like to augment their speech with on-the-fly visuals, we conducted two brainstorming sessions. These results informed a design space for visually augmenting conversations.

3.1 Procedure

We recruited 10 participants via group email invitations and internal communication channels in Google. Participants had various technical and non-technical backgrounds, including software engineers, researchers, UX designers, visual artists, students, etc. We held one-hour brainstorming sessions with two groups of five participants. In each session, we introduced the low-fidelity prototypes of the envisioned system, followed by video demos of the existing text-to-image systems. Participants then brainstormed ideas on a digital white board based on three prompts: (1) What are some scenarios where you could imagine using real-time visuals to augment conversations? (2) What types of visuals would you like to add in conversations? (3) How would you like to present and interact with the visuals?

3.2 Design Space for Augmenting Verbal Communication with Visuals

Two researchers organized participants' responses with the affinity diagram approach. Informed by the set of low-level and high-level themes derived, we developed a design space for systems that augment verbal communication with visuals. We followed the design space analysis methods [6] and held iterative discussion sessions. We identified eight key dimensions as detailed in Figure 2.

D1. Temporal. To augment verbal communication with visuals, systems can be either *synchronous* or *asynchronous*. The majority of prior systems provides augmentations asynchronously. Users either have to set up corresponding visuals before (*e.g.*, pre-configure visuals for an upcoming presentation, and trigger visuals by gestures and keywords [42]), or select and edit visuals after the text is composed [25, 51]. Our system falls in the paradigm of synchronous augmentation, where users select appropriate visuals on-the-fly while engaging in conversations.

D2. Subject. Visual augmentations of spoken language can either be used by the speaker to express their ideas (visualize their own speech) or by the listener to understand others (visualize others' speech). The majority of prior art in this domain falls under the former paradigm, where speakers select and design corresponding visuals to support their speech. In our system, we aim to support both subjects and allow all parties to visually supplement their own speech and ideas.

D3. Visual. Participants in our formative study wish to augment speech using a variety of visuals. We identified three main aspects to consider when providing visual augmentations:

(1) Visual Content — what information to be visualized? A segment of speech contains different information that can be visualized. For example, consider the statement *"I went to Disneyland with my family last weekend"*. One could visualize the generic term *Disneyland*, a picture of *I*, or more specific, contextual information such as *me and my family at Disneyland*. The system should be able to

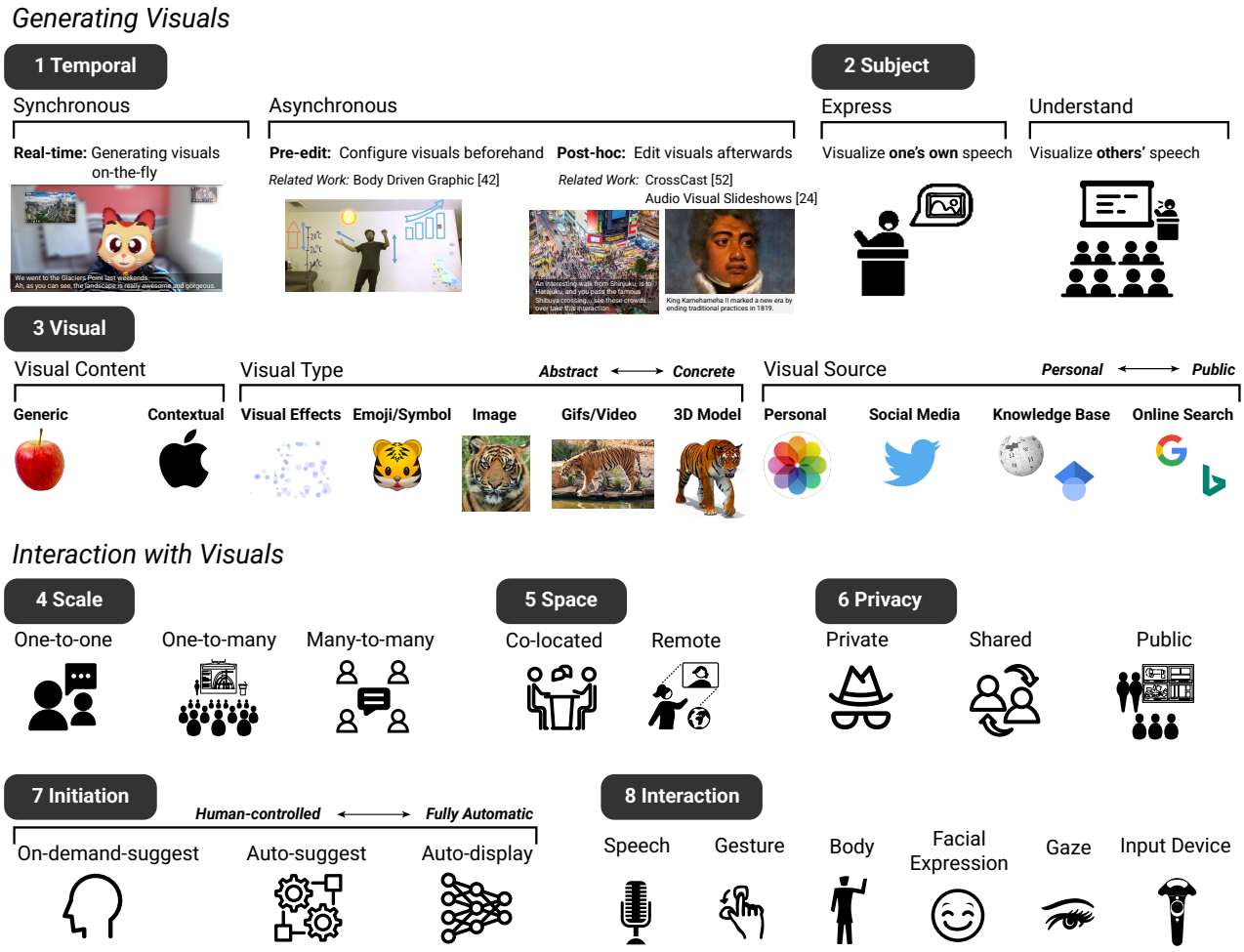

Figure 2: Design space for augmenting verbal communication with visuals. Visual Captions focuses on generating synchronous captions with a wide range of visual content, type, and sources. Our lab study allows users to express oneself in one-to-one remote conversations, while we deploy Visual Captions in the wild to explore one-to-many (when a single individual sends messages to a large audience, *e.g.*, giving a presentation) and many-to-many (when multiple people communicate with each other simultaneously, *e.g.*, group social event) scenarios. Visual Captions also provides three levels of AI proactivity to accommodate different interaction preferences.

disambiguate the most critical and relevant information to visualize in the current context.

(2) Visual Type — how should the visual be presented? There are often multiple ways to present a visual, ranging from abstract to concrete. For example, the term *Disneyland* could be visualized as: an icon of Disneyland, a photo of Disneyland, an interactive 3D map of Disneyland, or a video of people riding a roller-coaster. While visuals may have similar meaning, they can evoke different levels of attention and provide different levels of detail.

(3) Visual Source — where the visual should be retrieved from? Different sources can be utilized for the visual augmentations, including both personal and public assets. One might want to retrieve personal photos from one's own phone, or public images from the internet. While personal photos provide more contextual and specific information, images from the internet can provide more generic information with less privacy concerns.

Our system leverages a large language model optimized to consider the context of conversations, and identify the most appropriate visual content, type and source to suggest.

D4. Scale & D5. Space. Visual augmentations could be used in various communication scenarios, including one-on-one meetings, one-to-many lectures and many-to-many discussions. The number of participants and their location (*e.g.* in-person *v.s.* remote) can greatly affect best practices for such visual augmentations. We developed Visual Captions for existing video conferencing software

Use Cases & Scenarios	#P
Education and Lectures • Math&Sciences: *"A math teacher demonstrates an orientational relationship between objects."* • History: *"History teachers could present Napolean when it comes to famous battles in Waterloo."* • Language Learning: *"Visualize unfamiliar words when learning a new language"*	10
Casual Chats • Introducing Pets: *"When talking about my dog show a picture for introduction."* • Unknown Dishes: *"When Ordering food at a Japanese Restaurant, see what a Sukiyaki is."* • Movie Information: *"Show movie posters when talking about recent movies."*	10
Business and Utility • Visual Navigation: *"Show directions with images of buildings."* • People's Names: *"Visualize people's profile in meetings."* • Mind Map Search: *"Have idea of something in mind but forgot - work with AI to find right imagery."*	8
Creativity • Project Brainstorming: *"Provide immediate visual materials for brainstorming to spur creativity."* • New Topics: *"Expand upon visuals to find new topics."*	5
Storytelling • 3D Animals: *"When telling a story to children, display 3D models of animals"*	3

How Real-time Visuals Help		#P
Information	Provide and receive more information and new knowledge	10
Clarification	Resolve ambiguity and misunderstandings	10
Search	Quickly retrieve visuals from online search or photo album	8
Engagement	Make verbal communication richer and more interesting	6
Accessibility	Easier to understand for people with disabilities / with language barriers	5
Inspiration	Provide new topics and move the conversation forward	5
Memorization	Better memoriza content based on visuals	2

Figure 3: From our formative study, participants reported a number of use cases of Visual Captions and how visuals would facilitate communication. We list the reported themes and the number of participants who mentioned the theme (#P).

to augment meetings at different scales, supporting one-on-one, one-to-many, many-to-many scenarios.

D6. Privacy. Visual augmentations should take privacy into consideration right at the beginning and allow users to select among multiple privacy options: 1) Privately shown visuals are only presented to the speaker, invisible to any audience. 2) Publicly shown visuals are presented to everyone in the conversation. 3) In-between, the visuals can be selectively presented to a subset of audiences. We provide speakers with options 1) and 2) and to privately preview the visualizations before displaying them to the audiences. Listeners could also use our system to privately see the visuals based on speech they hear.

D7. Initiation. 6 participants in our formative study wanted the least efforts to generate visuals during the conversation, and therefore prefer having the system proactively providing visual augmentations without user interaction. However, other participants would like to have more control over the visuals, including when to trigger them and what to show. To meet these different preferences, we designed three levels of AI proactivity: on-demand-suggest, auto-suggest, and auto-display.

D8. Interaction Participants mentioned six domains of potential interactions: speech (*e.g.*, *"let's show an image here"*), gesture (*e.g.*, pinch), body pose (*e.g.*, waving hands), facial expression (*e.g.*, to trigger emojis), gaze (*e.g.*, for selecting visuals from suggestions), and custom input devices (*e.g.*, a controller). We support traditional input devices (*e.g.*, keyboard, mouse, and touch screens) given their universal use in video conferencing. In Visual Captions, we trigger visual generation by understanding the language via speech-to-text engines or by user pressing the space bar. In the future, we can expand capabilities to enable interaction with body pose, facial expressions, and other devices.

During the session, participants discussed various potential use cases for Visual Captions and how they believe it would be helpful (Figure 3). Many participants expressed a desire to use Visual Captions for educational and casual scenarios, and said that they would appreciate the addition of visuals to make conversations more informative and clear.

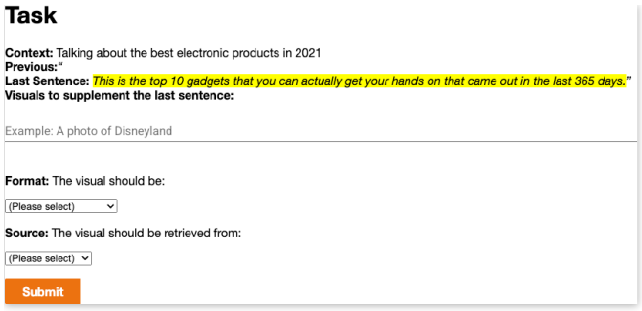

Figure 4: Crowdsourcing interface on MTurk we designed to collect dataset about people's preferences in augmenting verbal communication with visuals. In the crowdsourcing task, participants were asked to suggest visuals to supplement the *last sentence* in the conversations. We also show the previous sentence under *Previous*.

4 VC1.5K DATASET

To further understand how people would prefer to augment their verbal communication with visuals, or their *visual intents*, we collected a dataset of 1595 sentence-visual pairs across a wide range of contexts. The VC1.5K dataset is available at: https://liubruce.me/visual_captions.

4.1 Data Collection

We collected a total of 1922 sentences (transcribed speech) from 42 YouTube videos (1201 sentences) and the DailyDialog dataset [26] (721 sentences). We manually categorized each video source and

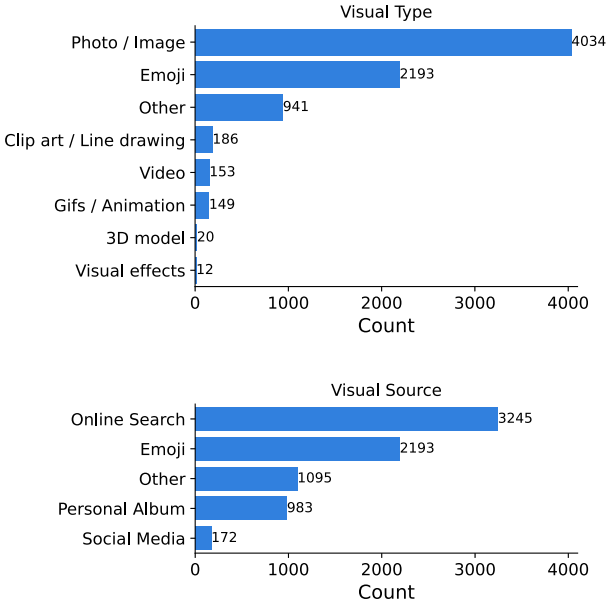

Figure 5: Distribution of the crowdworkers' responses of the appropriate visual sources and types.

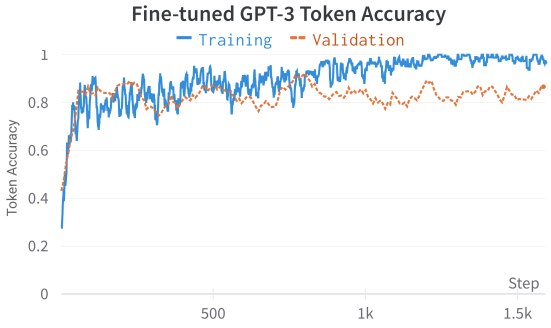

Figure 6: Model performance over 1500 steps.

added a brief description of its context (*e.g., best electronic products in 2021*). Videos were selected from various topics (Figure 15) ranging from tour guide, lectures to documentaries etc. For each video, we downloaded its automatic captions and segmented the transcriptions into sentences using NLTK's Sentence Tokenizer[2]. Our dataset has few errors in sentence segmentation because all captions contain punctuation. We observed some short sentences (*e.g. "Pop!"*), and mitigate this by also providing the previous sentence in both the labeling and training process.

We then crowdsourced visual intents through Amazon Mechanical Turk (MTurk). We provided specific task instruction and examples to crowdworkers before they start the task (Figure 13). For each annotation, we provided a short description of the video context, as well as the previous two sentences of the conversation. The crowdworkers were asked to *"Detemine what visual content could be shown to supplement the last sentence, given context and previous conversation"*, and write their answer in the format of *Visual Type* of *Information To Visualize*. We additionally asked crowdworkers to select an appropriate visual type and visual source from a list of categories we identified from our formative study (Figure 4). Crowdworkers were allowed to submit multiple visuals per sentence. If there was no relevant visual content, they were asked to submit "none". Each sentence was annotated by 4 different crowdworkers, and workers were allowed to work on multiple HITs. We selected workers to be within the United States and have a history approval rate beyond 95%. We paid crowdworkers $0.15 for each HIT completed. A total of 246 unique crowdworkers were employed.

We post-processed the raw data to ensure that the annotations were in the correct format and that the crowdworker responses were consistent. Specifically, we checked that the visual type mentioned in the answers matched the visual type selected by the worker. We rejected and republished any tasks that did not pass these quality checks.

As our goal is to obtain a ground truth for what people want to show visually, we combined crowdworker responses by sentence similarity. Specifically, we first removed all the English stop-words and punctuations, and used NLTK's WordNetLemmatizer[3] to lemmatize the remaining words and transformed all the words to lowercase. We then used SentenceTransformers [38] to compute two sentences' similarity. We consider two visual intents to be the same if their similarity score is greater than 0.85. After merging, 1595 sentences in our dataset have visual intents agreed by at least two crowdworkers, and 682 sentences by at least three crowdworkers. We removed all the sentences with less than two agreed visual intents. The final dataset (VC1.5K) consists of 1595 examples. We visualize the distribution of visual sources (Figure 5), and visual types (Figure 5) crowdworkers annotated.

5 MODEL

We trained a visual intent prediction model that predicts what visuals would be appropriate to show to supplement the ongoing conversation. After experimenting with different methods, including named entity detection, emotion analysis, and knowledge graph search, we used GPT-3 [5] fine-tuned with our VC1.5K dataset. This allows for a contextual system that can work on a range of topics without needing to be trained on specific vocabulary. We describe our training process, empirical analysis of model capabilities, and a technical evaluation.

5.1 Visual Intent Prediction Model

For training, we parsed each response into the format of "`<Visual Type>` of `<Visual Content>` from `<Visual Source>`". This serves as the label of our model. If there are more than one eligible label, we concatenate all responses together with a separator "`;`". If crowdworkers did not annotate any visuals for the sentence, we label "none". The prompt for each example is set to the previous two sentences in the text data. We additionally added a fixed separator,

[2]NLTK's Sentence Tokenizer: https://www.nltk.org/api/nltk.tokenize.html

[3]https://www.nltk.org/_modules/nltk/stem/wordnet.html

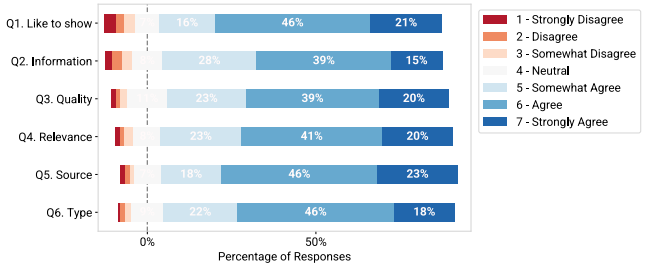

Figure 7: Technical evaluation results of the visual prediction model rated by crowdworkers.

"→", to inform the model when the prompt ends and the completion begins; and a fixed stop sequence, "\n", to indicate when the completion for the current example ends. In summary, we template the training data as:

{"prompt": "<Previous Two Sentences> →",
"completion":
"<Visual Type 1> of <Visual Content 1> from
<Visual Source 1>;
<Visual Type 2> of <Visual Content 2> from
<Visual Source 2>;
...
$\backslash n$"}

We used 1276 (80%) examples for training and validation, the remaining 319 (20%) examples as test data. Data in each of the training, validation and test sets were selected from different videos to prevent overlaps. We fine-tuned the `text-davinci-002` model[4] for 4 epochs with a batch size of 8, at a learning rate of 0.05. API access to our model is available at https://liubruce.me/visual_captions.

We measured the performance of the fine-tuned model with the *token accuracy* metric, *i.e.*, the percentage of tokens in a batch that were correctly predicted by the model. In the process of training we did not perform any hyper-parameter optimization. During training, our model reached a training token accuracy of 96.81%, and a validation token accuracy of 86.59% (Figure 6). We additionally evaluated the categorical prediction accuracy of visual type and visual source on the test set. Our model reached an 87.6% accuracy for visual type and an 86.1% accuracy for visual source predictions.

5.2 Technical Evaluation

To better understand the generality of our model and the quality of visual suggestions, we conducted a technical evaluation with crowdworkers recruited on MTurk. We first generated visuals for the out-of-bag test dataset using our model. 282 visuals were generated from the 266 (out of 319 sentences in total) examples that had at least one visual suggestion (*i.e.*, not "none"). In this evaluation, we retrieved all visuals online, and provided crowdworkers with the original sentence, predicted visual content, visual type, and visual source. We asked crowdworkers to rate the visual suggestions on four dimensions using a 7-point Likert-scale from *1–Strongly Disagree* to *7–Strongly Agree*. Questions are listed in Appendix subsection C.1 Each example was evaluated by three crowdworkers (a total of 846 tasks). As an attention test, we also asked crowdworkers to briefly explain their reason for the first question. We examined crowdworkers' responses, rejected and republished 139 low quality results that were too short, ambiguous, or irrelevant. We selected workers with approval rates over 95% and a US-based location, and paid them $0.30 per task completed. A total of 89 workers were employed.

Overall, our model and visual suggestion pipeline produced desirable visuals (Figure 7). Crowdworkers rated the statement *"I would like to display the visual when having the conversation (Q1)"* with a greater than or equal to *5–Somewhat Agree* 83% of the time. They consider the displayed visuals to be useful and informative (Q2, 82% ≥ *5–Somewhat Agree*), high-quality (Q3, 82% ≥ *5–Somewhat Agree*), and relevant to the original speech (Q4, 84% ≥ *5–Somewhat Agree*). In addition, crowdworkers found the predicted visual type (Q5, 87% ≥ *5–Somewhat Agree*) and visual source (Q6, 86% ≥ *5–Somewhat Agree*) to be accurate given the context of the corresponding conversation.

5.3 Visual Suggestion Examples

We analyzed a subset of our visual intent prediction model's results on the test set and qualitatively showcase some examples in Figure 8. The model demonstrates good performance and consistency across different topics. It understands the context of a sentence and recommend different visual types, content, and sources based on that context. It also suggests multiple visuals if there are multiple visual intents in a sentence, and can retrieve accurate visuals for ambiguous descriptions. These behaviors were encoded in our collected dataset and training process. In our test dataset, 27 out of 319 sentences have multiple visual suggestions. In future, we would explore training our model on a wider range of visual intents, estimating the importance/appropriateness scores of visuals and allowing users to filter for more or less visual suggestions.

6 VISUAL CAPTIONS SYSTEM

With the fine-tuned visual intent prediction model, we developed Visual Captions on *ARChat* (details in appendix subsection A.1), a web-based rapid prototyping platform 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.

6.1 Visual Captions Interface

6.1.1 Generating Visual Suggestions. Visual Captions automatically suggests relevant visuals based on users' conversation content (Figure 9). 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 (6.2). 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

[4]OpenAI's Fine-tuning API: https://beta.openai.com/docs/guides/fine-tuning

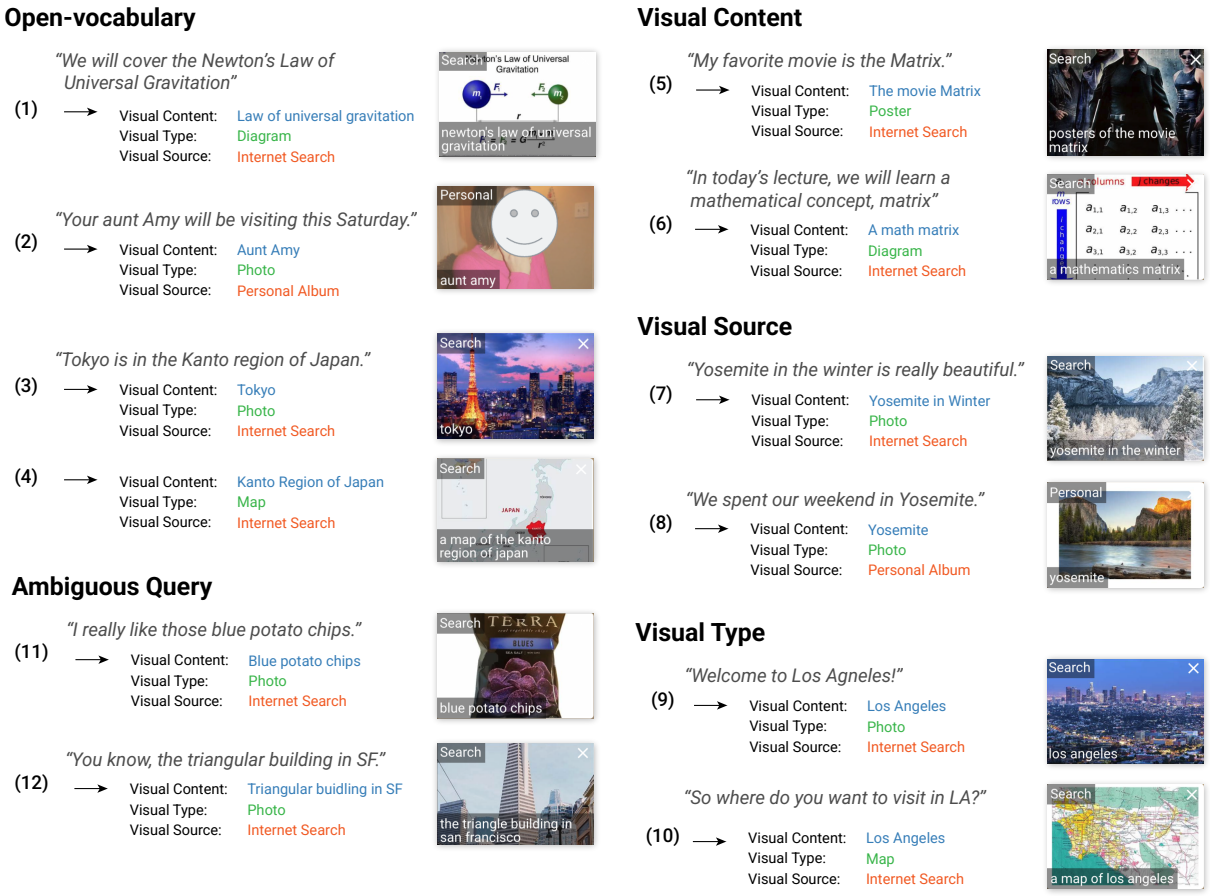

Figure 8: Examples visual intent predictions by our model. Compared with keyword-based approaches, our system could handle open-vocabulary conversations and contextually predict visual content, visual source, and visual type.

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 [36], *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"*, using the Microsoft Bing Image Search API[5]

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 10), 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.

[5]Bing Image Search API: https://microsoft.com/bing/apis/bing-image-search-api

6.1.2 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 private *Scrolling View* (Figure 10A). The scrolling view automatically updates when new visuals are suggested by the system, and removes the oldest visual widget if it exceeds the maximum amount (customizable *# Max Visuals* in the settings page). Similarly, emojis are displayed in a separate scrolling view in the bottom right corner of the screen (Figure 10C). When proactively suggesting visuals, our system on average suggests a new visual every 7.6 seconds in our user studies.

6.1.3 Spotlight View. To make an image visible to all parties, the user may click the widget to move it to the *Spotlight View* (Figure 10B). Visuals in the spotlight view can be moved, resized, and deleted.

6.2 Visual Captions Settings

The Visual Captions settings page (Figure 14) 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

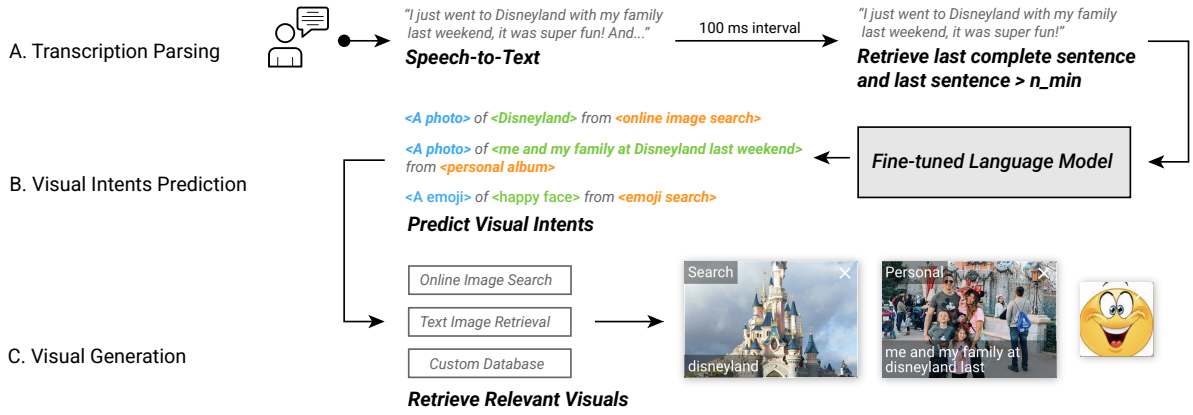

Figure 9: System workflow of Visual Captions. The workflow consists of three major steps: A. Transcription Parsing; B. Visual Intents Prediction; C. Visual Generation.

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 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.

6.3 AI Proactivity in Visual Captions

Informed by our formative study and pilot study, 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.

7 EVALUATION

We first conducted a user study with 26 participants to evaluate Visual Captions. We used a mixed-methods study design to gather both data from participants' survey responses and semi-structured interviews. The study examined the following two research questions:

RQ1: How do people use visuals 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 human-human activities?

7.1 Pilot Study

We conducted three one-hour long pilot study with 6 participants (3 pairs) to gather initial feedback around the interaction and interface design of Visual Captions. All 6 participants were recruited within Google. In the pilot study, participants (labeled as P1 to P6) were asked to use Visual Captions for a scripted conversation and an open-ended conversation to test out the system, followed with a contextual interview around system's usability.

Participants identified three main issues: the emoji suggestions were distracting and less useful, the lack of an "on-demand" mode, and the lack of customization options. In response, Visual Captions was updated to move the emoji suggestions to a separate, ambient area of the screen and reduce their size, add an "on-demand-suggest" mode which only suggests visuals when the user explicitly requests them, and add a settings page to allow for customization. We integrated all improvements and conducted our formal study with an additional 20 participants.

7.2 Materials

We created three scripted dialogues on different topics: discuss where to visit in Los Angeles, order food in a Japanese restaurant, and chat about what they did over the weekend. All conversations were around 5 minutes. Conversations were scripted to provide controls for our data analysis, and allow Visual Captions to suggest a variety of visuals for both participants throughout the study. Participants were also allowed to go off-topic a little bit during the scripted conversations.

7.3 Participants

We recruited 20 participants (9 female and 11 male) from our institution's email list and internal communication channel, labeled as P7 to P26. Participants were 21–61 years old (avg=29.5, std=9.4). Participants had varying technical and non-technical backgrounds including students, software engineers, research scientists, designers, and product managers etc. 10 participants use video conferences multiple times per day, five daily, four multiple times per week, and

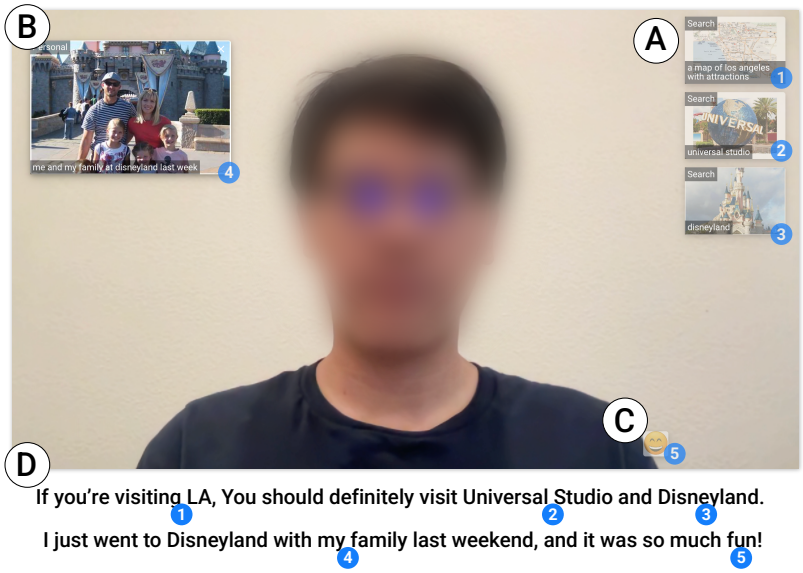

Figure 10: In Visual Captions's interface (default auto-suggest mode), the *scrolling view* (A) displays privately candidates of visual suggestions generated by our visual intent prediction model. Emoji suggestions are displayed on the bottom right corner (C). Users can click and display the visual to the *spotlight view* (B) to share it publicly.

one weekly. Participants were recruited in pairs to fit the one-on-one conversation scenarios in our study. Eight participants were not close friends with their conversation partner. Each participant was compensated with a $25 gift card for their completion of the study.

7.4 Procedure

We conducted a one-hour-long study with each participant pair. We set up two meeting rooms in our office. Each meeting room was equipped with a laptop computer connected to a 32-inch monitor, an external keyboard, and a wireless mouse. We created a Google Meet room with Visual Captions connected before the study. We also spent 10 minutes instructing participants how to use Visual Captions and its settings page. Each participant tried all features and experimented with the three AI proactivity levels before continuing to the main part of the study.

Each participant pair was asked to act out four scripted conversations and have a 5-10 minutes open-ended conversation with each other. We divided the study into three parts: (1) Verbal communication *v.s.* communication with visuals. First, participants act out the same scripted conversation two times to directly compare their experience of conversation with and without visuals. All participants used the auto-suggest mode. (2) AI proactivity. Participants then proceed to test the auto-display and on-demand-suggest mode with two additional conversations. (3) Open-ended conversation. Participants use their preferred customized settings to open-endedly chat with their study partner. After each conversation, we recorded participants Task Load Index [14] and Likert scale ratings on their feeling of control, intrusion, error, and interruption when interacting with the system. After the open-ended conversation, participants rated the usefulness of visuals in conversations (questions in appendix subsection C.2). Lastly we asked participants to compare their overall experience with and without visuals and different AI proactivity levels through semi-structured interviews. We recorded the audio track for the entire study and the screen portion of using the interface.

7.5 Findings: Using Real-time Visuals in Conversations (RQ1)

We present our findings on the impact of visual augmentations in conversations (RQ1) and how people prefer to interact with AIs with different proactivity levels (RQ2).

7.5.1 How do people select and display visuals? Participants in our study exhibited different practices when selecting and displaying visuals. In the auto-suggest mode, some participants only looked at visual suggestions in the scrolling view when they wanted to show something visually. For instance, P9 mostly focused on the conversations and would not notice visuals coming up. For him it was more of a *"purposeful action"*. In contrast, P9's study partner P10 continuously looked for new visuals whenever something popped up, and always *"take a quick look to see if it was something useful to show"*. P15 reported glancing at suggested visuals every time he finished his sentence.

14 participants described that they would click and display a visual to their study partner when they were trying to convey complementary information. P14 and P21 also showed visuals when they were trying to emphasize things. Participants' decision of whether to display a visual or not was largely impacted by the conversational context and social norm. For example, P11 would use more visuals in relaxed and casual conversations as something funny to chat about. P18 said *"it depends on who you're talking to"*.

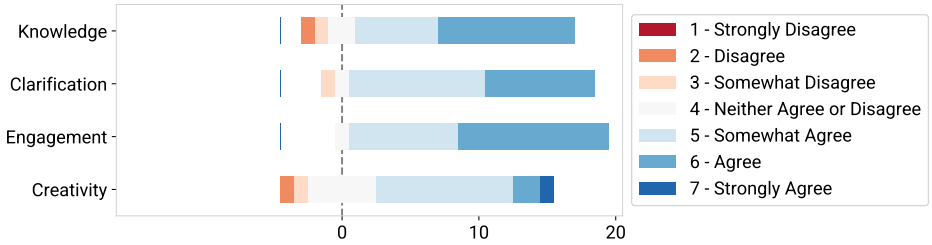

Figure 11: Participants' Likert scale ratings to (1) Knowledge: The suggested visuals can help me understand some unfamiliar terms and concepts; (2) Clarification: The suggested visuals can help clarify ambiguity in the conversation; (3) Engagement: The suggested visuals can make the conversation more interesting and enjoyable; (4) Creativity: The suggested visuals provided topics to talk about or led the direction of the conversation.

P13 mentioned her hesitation when the system suggested a visual based on a question she asked:

> *"Sometimes it shows visual answers of what I was asking for, for example 'what is tempura?'. Like should I be the one to show it?"* – P13

7.5.2 How are visuals useful? **Helps Explain and Understand Unfamiliar Concepts.** Participants found real-time visuals to be useful in conversations in multiple aspects. From the study survey responses, participants rated the statement "The suggested visuals can help me understand some unfamiliar terms and concepts and make the conversation more informative" with a *5—Somewhat Agree* or greater 80% of the time (Figure 11). They found Visual Captions useful to learn unfamiliar concepts or words without having to ask explicitly:

> *"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

In addition to understanding unfamiliar concepts in conversations, participants found Visual Captions to be helpful in explaining things to their conversational partners. P14 used visuals to more easily describe places to visit to her friend:

> *"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."* – P14

Makes Known Information More Intuitive 6 participants also reported that visuals helped them with known phrases by making the information more intuitive and easy to convey. For example, P20 enjoyed having visuals of famous paintings to supplement his description: *"When I talked about the Starry Night painting it was really good to have a picture there."* Participants reported that when they wanted to quickly explain a concept without having to describe all details, they found visuals helpful:

> *"Having it pop up and being able to see what the dishes would have been extremely useful. We just had a visual representation of the appetizers, and be like oh I want this one and that one, and then we would see it very quickly."* – P16

Clarifies Ambiguities in Language. Participants reported that Visual Captions helped them clarify ambiguities in their verbal conversations. In several cases, participants used visuals to clarify objects or information that was abstract. 18 participants (Figure 11) agreed that the suggested visuals help clarify ambiguity in conversations (≥ *5 — Somewhat Agree*). P21 mentioned that during her conversation, *"whenever you're bringing up something that he wasn't sure which I'm talking about, you can just select the right image."* Similarly P17 described a case where he could select the correct visuals to illustrate what he meant:

> *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.* – P17

Makes Conversations More Fun and Engaging. 19 participants (Figure 11) found visuals to help make conversations more fun and engaging (≥ *5—Somewhat Agree*). Specifically, participants were able to communicate with more information and *"it makes the conversation longer and more interactive"* (P9). Participants reported that they had richer conversations with their conversation partners:

> *"When chatting with recommended pictures, our interaction has increased. The conversation is getting longer with more content."* – P14

Interestingly, four participants reported that sometimes incorrect visuals suggestions also made their conversations more fun: *"Especially when it suggests funny visuals that were not expected. Sometimes the AI errors were kinda funny."* (P12)

Reduces Efforts in Search for Visuals. Participants found that Visual Captions reduced laborious search and sharing screen when explaining certain concepts: *"It is very interesting and necessary to have them. If we talk about something, we search and this is much faster and easier"* (P11). Two participants mentioned they could grant the system access to their personal albums and Instagram posts when talking to their parents or close friends. P20 said it would be *"so much easier for me to show photos from my recent trips when talking about it"*.

7.5.3 How visuals impact the way people communicate? **Visual Captions Guided Conversations and Provided New Topics.** In open-ended conversations, participants found that Visual Captions

guided their conversations as they would often start talking about the visuals when they showed up. For example, P13 described:

> *"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."* – P13

Visuals are also helpful to find new topics to get to know each other, especially for those who were unfamiliar with their study partners. 13 participants rated the statement "The suggested visuals provided topics to talk about or led the direction of the conversation" with greater or equal to *5—Somewhat Agree* (Figure 11). Moreover, in our scripted conversations, a majority of the participants did not strictly follow the script and sparked emotional remarks or engaging actions. For example, after P11 said *"Let me show you a picture of Rodeo Drive"*, P12 answered *"Wow, it appears looking really pretty"* instead of the *"I'll have to check it out"* as originally scripted.

Distractions of Visual Captions. To our surprise, participants did not find conversations with extra visuals to be more mentally (Mann-Whitney U test $p > 0.05$) or physically demanding ($p > 0.05$) than normal conversations when the system is proactively suggesting visuals (*i.e.* auto-display and auto-suggest), and did not find it to be more stressful or annoying ($p > 0.05$) in any condition (Figure 12). While many participants found Visual Captions to be *"not distracting at all"* (P21), some found Visual Captions to disrupt the conversation flow in two ways: interaction required for selecting and displaying visuals (more discussion in subsection 7.6), and deciding what visuals to display. Participants reported that sometimes they needed to switch their mind to decide whether a visual is appropriate:

> *"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

7.5.4 Improvements and Opportunities. Over half of the participants envisioned to use Visual Captions with augmented reality glasses. Participants suggested that it would be more natural and less intrusive, especially for in-person conversations. P12 particularly mentioned the use of hand gestures to intuitively select and display visuals. Other participants also mentioned accessibility use cases of Visual Captions for people with dyslexia or language barriers. In addition, many participants were excited about expanding this tool into professional settings, and use it in their work to retrieve slides, profiles, internal links etc.

7.6 Findings: Interacting with AI (RQ2)

7.6.1 People have a large variance of AI proactivity level preferred. Participants in our user study had diverging preferences over the AI's proactivity level, even within the same study pairs. Six participants preferred the *auto-display* mode, seven preferred the *auto-suggest* mode, and seven preferred the *on-demand-suggest* mode.

Auto-Display Mode. Participants who preferred the auto-display mode liked that almost no interaction was needed to activate and display the visuals, as P7 mentioned, *"Not having to click is huge for me."* On a scale from 1 to 7, participants rated their effort using the auto-display mode ($\mu = 1.35, \sigma = 0.73$) to be less than auto-suggest ($\mu = 4.05, \sigma = 1.07$) and on-demand-suggest modes ($\mu = 4.4, \sigma = 1.28$), with statistical significance (Mann-Whitney U test $p < 0.0001$).

However, some participants complained about it being prone to errors and having less control: *"Sometimes I don't want to emphasize the point, but it suddenly shows an image..."* (P9). From our survey responses, participants felt that the auto-display mode provided significantly less ($p < 0.01$) control ($\mu = 2.55, \sigma = 1.32$) compared to auto-suggest ($\mu = 4.9, \sigma = 1.67$) and on-demand-suggest ($\mu = 5.15, \sigma = 1.28$), and were harder to override system mistakes ($p < 0.001$). For example, P17 had to pay attention all the time to not risk showing something inappropriate: *"It interrupt my conversation because I always pay attention to the spotlight view, and immediately check whether the image displayed is appropriate."* In contrast, participants who preferred auto-display felt that the incorrect visuals did not matter that much:

> *"If it shows an image incorrectly, I just feel like it's an imperfect tool. Just like textual captions, it's not accurate and I just accept that fact."* – P20

Privacy is another concern in the auto-display mode. Although we did not observe any statistical difference in participants' responses, some participants mentioned that the system could accidentally show private photos without permission: *"I don't want to show my personal album without confirmation."* (P7) In our user study, while our system can suggest personal visual intents, we did not use real images from personal albums for privacy reasons. This is a limitation of the study, and it would be important to understand how people would feel about having real personal visuals suggested or displayed, especially in the auto-display mode. In future, implementing safeguards such as an extra confirmation of private photos may give users more control over their privacy.

Auto-Suggest Mode. Seven participants preferred the *auto-suggest* mode, as they felt that they could better understand the system's capabilities in the private scrolling view and share with others when needed: *"During the conversation, I really like talking when these things pop out, so we really know what it is like"* (P8). Participants who preferred this mode usually *"only look for visuals when I want to show something."* (P12).

Other participants did not like selecting from the scrolling view: *"It takes some time to identify which images are proper to share."* (P22). Some mentioned that this mode increased the cognitive burden and distracted their attention as visuals are being constantly suggested in the scrolling view:

> *"For me it was kind of a distraction from the the conversation because typically I don't look away from the person when I talk, and I think having something moving in the frame is kind of distracting"* – P14

On-Demand-Suggest Mode. Although participants rated the on-demand-suggest mode to the most mentally and physically demanding (Figure 12), seven participants preferred this mode. Participants considered it to be the least distracting and appreciated having full control over when to query for a visual:

> *"I prefer the tap one since I feel like I was controlling whether I want to show something. It's less mental overload and distraction because I would only activate it*

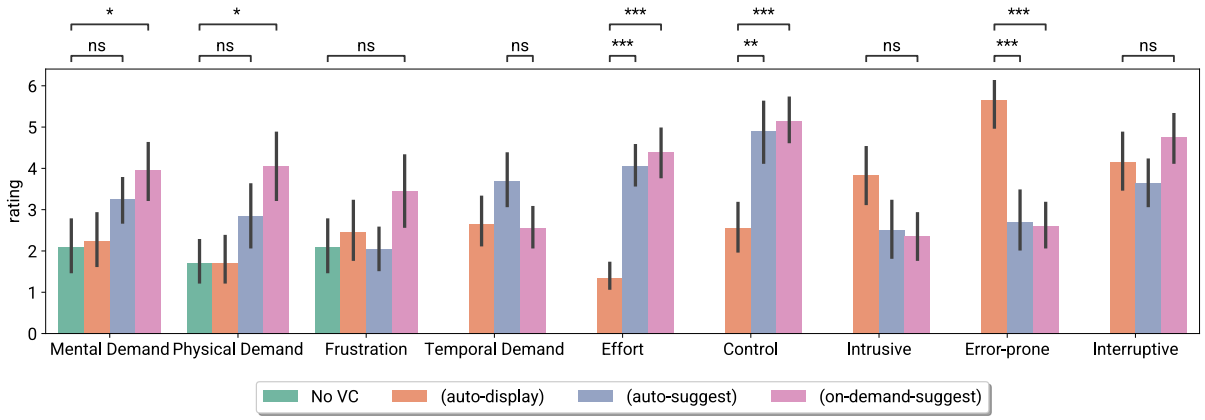

Figure 12: Participants' Task Load Index and Likert scale ratings (from 1 - Strongly Disagree to 7 - Strongly Agree) to four conversations with different Visual Captions modes: No VC, Auto-Display, Auto-Suggest and On-Demand Suggest, with 95% confidence interval bar. In the Task Load Index questions we asked participants their feeling of mental demand, physical demand, frustration (how stressed and annoyed), temporal demand (how hurried and rushed), and effort. In Likert scale questions we asked participants about their experience interacting with the AI, and their feeling of having control, AI intrusiveness, difficulty to override errors, and AI's interruption to the conversation. We additionally annotated statistical tests (Mann-Whitney U) results between different conditions: ns (non-significant), * (p<.05), ** (p<.01), * (p<.001).**

> *when I want, and <it> also <has> less things on the screen."* – P13

P17 also found this mode to seemingly provide the most accurate visual suggestions, because *"I'm only seeing visuals I requested."*

Participants who did not like on-demand-suggest felt too much interaction effort was needed to display a visual. Interestingly, although the latency of visual suggestions should be the same for all three modes on average, participants repeatedly reported that they felt significantly more delay when using the on-demand-suggest mode. We hypothesize that when humans are taking the initiative, people are more sensitive to when they started the query, and thus a more obvious feeling of delay. While with proactive AI, people are mostly paying attention to the main conversation. Because of this perceived delay, several participants found it hard to understand the internal state of the system: *"It's hard for me to know when to press the spacebar, and once I pressed the spacebar, did it trigger the system or there's just no suggestions?"* (P8). Participants sometimes had to interrupt their conversation to wait for something to show up:

> *"I have to pause for a bit after I hit the space and hoping that the right thing will show up. "* – P21

8 CASE STUDY: USAGE OF VISUAL CAPTIONS IN-THE-WILD

To further understand how people would integrate real-time visual augmentations into their daily conversations, we distributed Visual Captions let people try it in-the-wild for two weeks. We recruited 10 participants (5 females, age: 22 – 33, avg=28.5, std=3.8) via email invitations and social channels in Google, labeled as U1 to U10. Six users previously participated in our lab study. We encouraged participants to use Visual Captions in every meeting during the week. Participants filled in an entry survey and an experience sampling survey every three days.

Using Visual Captions Beyond One-on-one Conversations. One limitation of our case study is we recruited participants within our institution due to confidentiality, and participants had mostly work-related meetings where Visual Captions was not specifically designed for. However, all participants still used Visual Captions in various scenarios: 4 users used Visual Captions in multiple meetings per day, 2 users about once a day, and 4 users multiple times a week.

Participants reported not only using Visual Captions in one-on-one meetings, but also in team meetings, remote social events, and listening to online talks, where multiple parties were involved. *"I used Visual Caption in a team social with my colleagues for 45 minutes, we were talking about favorite board games and weekend plans."* (U7), *"I used VC with other team members to talk about their experiences with a demo. Emotions during the talk were conveyed by emojis."* (U3). In such many-to-many scenarios, participants found Visual Captions especially attract attention from the audience:

> *"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."* – U2

Visual Captions Improve Expressiveness and Reduce Social Awkwardness. Some participants used Visual Captions as an implicit self-expressive tool for listeners. They noted that Visual Captions was especially helpful in large team meetings, as it allowed them to express their thoughts and feelings without interrupting others, through a parallel visual channel. *"Showing the suggested emoji on the screen also prevents interrupting with other people's talking while expressing oneself's agreement or rejection."* (U5)

In addition, Visual Captions reduces embarrassment and frictions for foreigners or laymen of specific topics in social scenarios. For non-native speakers, Visual Captions offers a parallel channels for

users to quickly catch up with concepts that are familiar to others. "*It helped understand unfamiliar words in English as a non-native speaker. E.g., Andromeda.*" (U3) "*'VC pops up images for words that I don't understand, like 'groomhaven', 'borg sphere' in a social meetup.*" (U9)

Users Prefer Different AI Proactivity in Different Scenarios. All six participants who participated in our lab study reported continuing using their preferred AI proactivity level in the case study. However, many participants reported that during the deployment, they were more likely to change the AI proactivity levels depending on the type of meetings, indicating their preferences for different levels of AI proactivity in different social scenarios. For example:

> *"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."* – U6

In one-on-one meetings, users may prefer a higher level of AI proactivity to enhance their own expressiveness, while in important group meetings they may prefer a lower level of AI proactivity to avoid interrupting other speakers. Similarly, U7 described that *"I used VC to show images of my own words in 1:1 meetings but show images of everyone's words in a group meeting using the auto-suggest mode."* This highlight the importance of providing users with flexible control over the level of AI proactivity in Visual Captions.

9 DISCUSSION AND FUTURE WORK

We describe the limitations of our system, discuss the implications and envision future opportunities:

Visual Appropriateness in Conversation. We identified two major causes of errors in Visual Captions. First, we observed that many of the visual suggestion errors come from incorrect captions. As our system relies on automatic speech recognition (ASR) as the input to the model, imprecise ASR results may lead to drastically incorrect visual suggestions. In future, we could leverage methods like topic modeling [48] and summarization [13] to compute a *visual appropriateness score*, to determine how relevant the visual is to the current conversation.

Several participants also mentioned that one of their main sources of distraction was to think about whether they want to display the visual. To this end, another aspect of the visual appropriateness measure would be to understand how appropriate it is to show the visual to others, given the current context, relationship with the person, and timing. Having a threshold to filter out all inappropriate visuals would be especially helpful in the auto-display mode.

Promptness-Accuracy Trade-off. The second main cause of error in Visual Captions is that it greedily retrieve unfinished sentences to ensure *promptness* in visual suggestions. For example, in the sentence *"San Francisco is a beautiful city"*, if the algorithm acts too aggressive it may incorrectly query *"San"*, leading to incorrect images. This problem of *when to query for visuals* is fundamentally similar to the problem of *when to transcribe* and *when to translate* in live speech. Recent research have proposed the Wait-K policy and to predict stability scores of new phrases [4]. In future we could employ a similar approach to search for visuals that have high probability to be stable.

Improvement of Visual Suggestions.

Personalized Visual Suggestions: Many participants use Visual Captions to explain information that their study partner did not know, and understand concepts that they were not familiar with. However, already known concepts may feel redundant. It still remains an open question that how to model people's familiarity with different concepts or knowledge, without privacy risks. In our case, we could explore online learning algorithms to continuously adapt the model to user's personal preferences.

Integrating Text-to-image Models: Recent advances in contrastive language models [36] and diffusion models in vision [18, 24, 41, 43] have fostered a few groundbreaking text-to-image systems: DALLE-2 [37], ImageGen [40], Parti [52], etc. While such models generate delicate visuals based on descriptions, our system infers the implicit visual intent of human conversations. However, it is interesting to explore integrating these text-to-image models into our system when people's visual intents are detected to be creative and imaginative, in scenarios such as brainstorming and storytelling.

Visual Coherence: It would also be beneficial for Visual Captions to consider coherence in visual styles and formats throughout a conversation. For example when discussing different data plots, it would be beneficial if the plots generated throughout the conversation were coherent in style, such as being generated with the same visualization tool like pyplot. In future, we plan to explore the use of consistent tools and style metrics to generate visuals that are coherent across a conversation.

The Future of Augmenting Verbal Communication with Visuals. In this paper, we explored augmenting verbal communication with visuals on videoconferencing platforms, as it is one of the most common way people communicate. However, as shown in our design space (subsection 3.2), we believe that the possible applications are much broader. For example, future systems could explore augmenting in-person conversations with 3D visuals and hand gestures in augmented reality or a metaverse like Geollery [11, 12], generating visual effects from verbal speech, or using real-time visuals for brainstorming and to assist creative writings.

Does a One-size-fits-all Human-AI Interaction Mode Exist? Participants had a large variance in the level of AI proactivity they preferred — 6 preferred auto-display, 7 preferred auto-suggest and 7 preferred on-demand-suggest. Participants also would like to use different AI proactivity levels under different social scenarios. While a number of recent research have proposed new human-AI collaborative systems and aimed to find an optimal human-AI interaction paradigm, there is an opportunity to investigate how to design and provide systems that are adaptive, with different levels of human and AI initiatives.

10 CONCLUSION

In this work, we presented our vision and implementation of Visual Captions, a human-AI system designed to visually augment real-time conversations. We shared how we worked with crowdworkers to gather 1595 visual intents to fine-tune a large language model to power our system, and report on the results from a formative study that informed the development of a design space for systems like ours. Our system, Visual Captions, was designed as a virtual camera and is therefore compatible with today's popular video

conferencing systems. As a Chrome browser plugin, it can leverage state-of-the-art real-time speech transcription and be easily installed by end users with no other dependencies. This approach allowed us to evaluate the capabilities in both a user study and a longer deployment study. The results suggest that Visual Captions has the potential to facilitate live conversations with valuable visual content.

ACKNOWLEDGMENTS

We would like to extend our thanks to Jason Mayes, Na Li, Yinda Zhang, Feitong Tan, and Ping Yu for participating weekly discussions, and our anonymous reviewers for their insightful feedback.

REFERENCES

[1] Salvatore Andolina, Valeria Orso, Hendrik Schneider, Khalil Klouche, Tuukka Ruotsalo, Luciano Gamberini, and Giulio Jacucci. 2018. Investigating Proactive Search Support in Conversations. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China) *(DIS '18)*. Association for Computing Machinery, New York, NY, USA, 1295–1307. https://doi.org/10.1145/3196709.3196734

[2] Salvatore Andolina, Valeria Orso, Hendrik Schneider, Khalil Klouche, Tuukka Ruotsalo, Luciano Gamberini, and Giulio Jacucci. 2018. SearchBot: Supporting Voice Conversations with Proactive Search. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Jersey City, NJ, USA) *(CSCW '18)*. Association for Computing Machinery, New York, NY, USA, 9–12. https://doi.org/10.1145/3272973.3272990

[3] Michael Argyle. 1972. Non-verbal communication in human social interaction. *Non-verbal communication* 2 (1972), 1.

[4] Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, and George Foster. 2020. Re-translation versus Streaming for Simultaneous Translation. In *Proceedings of the 17th International Conference on Spoken Language Translation*. Association for Computational Linguistics, Online, 220–227. https://doi.org/10.18653/v1/2020.iwslt-1.27

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models Are Few-Shot Learners. *Advances in Neural Information Processing Systems* 33 (2020), 1877–1901. https://doi.org/10.5555/3495724.3495883

[6] Stuart K Card, Jock D Mackinlay, and George G Robertson. 1991. A Morphological Analysis of the Design Space of Input Devices. *ACM Transactions on Information Systems (TOIS)* 9, 2 (1991), 99–122. https://doi.org/10.1145/123078.128726

[7] Jiajian Chen, Jun Xiao, and Yuli Gao. 2010. ISlideShow: a Content-Aware Slideshow System. In *Proceedings of the 15th International Conference on Intelligent User Interfaces* (Hong Kong, China) *(IUI '10)*. Association for Computing Machinery, New York, NY, USA, 293–296. https://doi.org/10.1145/1719970.1720014

[8] Peggy Chi, Nathan Frey, Katrina Panovich, and Irfan Essa. 2021. Automatic Instructional Video Creation From a Markdown-Formatted Tutorial. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 677–690. https://doi.org/10.1145/3472749.3474778

[9] Peggy Chi, Zheng Sun, Katrina Panovich, and Irfan Essa. 2020. Automatic Video Creation From a Web Page. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 279–292. https://doi.org/10.1145/3379337.3415814

[10] Pei-Yu Chi and Henry Lieberman. 2011. Intelligent Assistance for Conversational Storytelling Using Story Patterns. In *Proceedings of the 16th International Conference on Intelligent User Interfaces* (Palo Alto, CA, USA) *(IUI '11)*. Association for Computing Machinery, New York, NY, USA, 217–226. https://doi.org/10.1145/1943403.1943438

[11] Ruofei Du, David Li, and Amitabh Varshney. 2019. Geollery: A Mixed Reality Social Media Platform. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300915

[12] Ruofei Du, David Li, and Amitabh Varshney. 2019. Project Geollery.Com: Reconstructing A Live Mirrored World With Geotagged Social Media. In *The 24th International Conference on 3D Web Technology* (LA, CA, USA) *(Web3D '19)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3329714.3338126

[13] Yihong Gong and Xin Liu. 2001. Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New Orleans, Louisiana, USA) *(SIGIR '01)*. Association for Computing Machinery, New York, NY, USA, 19–25. https://doi.org/10.1145/383952.383955

[14] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*. Vol. 52. Elsevier, New York, NY, USA, 139–183.

[15] Kaveh Hassani and Won-Sook Lee. 2016. Visualizing Natural Language Descriptions: a Survey. *ACM Computing Surveys (CSUR)* 49, 1 (2016), 1–34. https://doi.org/10.1145/2932710

[16] Zhenyi He, Keru Wang, Brandon Feng, Ruofei Du, and Ken Perlin. 2021. GazeChat: Enhancing Virtual Conferences With Gaze-Aware 3D Photos. In *Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology (UIST)*. ACM, New York, NY, USA, 10. https://doi.org/10.1145/3472749.3474785

[17] Ilyena Hirskyj-Douglas, Anna Kantosalo, Andrés Monroy-Hernández, Joelle Zimmermann, Michael Nebeling, and Mar Gonzalez-Franco. 2020. Social AR: Reimagining and Interrogating the Role of Augmented Reality in Face to Face Social Interactions. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*. Association for Computing Machinery, New York, NY, USA, 457–465. https://doi.org/10.1145/3406865.3418585

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.

[19] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 159–166.

[20] Erzhen Hu, Md Aashikur Rahman Azim, and Seongkook Heo. 2022. FluidMeet: Enabling Frictionless Transitions Between In-Group, Between-Group, and Private Conversations During Virtual Breakout Meetings. In *CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3491102.3517558

[21] Yu Jiang, Jing Liu, Zechao Li, Changsheng Xu, and Hanqing Lu. 2012. Chat With Illustration: a Chat System With Visual Aids. In *Proceedings of the 4th International Conference on Internet Multimedia Computing and Service* (Wuhan, China) *(ICIMCS '12)*. Association for Computing Machinery, New York, NY, USA, 96–99. https://doi.org/10.1145/2382336.2382364

[22] Zhenhui Jiang and Izak Benbasat. 2007. The Effects of Presentation Formats and Task Complexity on Online Consumers' Product Understanding. *Mis Quarterly* 1 (2007), 475–500. https://doi.org/10.1016/j.dss.2015.03.001

[23] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 2019. Will You Accept an Imperfect Ai? Exploring Designs for Adjusting End-User Expectations of Ai Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300641

[24] Vladimir Kulikov, Shahar Yadin, Matan Kleiner, and Tomer Michaeli. 2022. SinDDM: A Single Image Denoising Diffusion Model. https://doi.org/10.48550/ARXIV.2211.16582

[25] Mackenzie Leake, Hijung Valentina Shin, Joy O. Kim, and Maneesh Agrawala. 2020. Generating Audio-Visual Slideshows From Text Articles Using Word Concreteness. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3313831.3376519

[26] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Asian Federation of Natural Language Processing, Taipei, Taiwan, 986–995. https://aclanthology.org/I17-1099

[27] Jian Liao, Adnan Karim, Shivesh Singh Jadon, Rubaiat Habib Kazi, and Ryo Suzuki. 2022. RealityTalk: Real-Time Speech-Driven Augmented Presentation for AR Live Storytelling. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 17, 12 pages. https://doi.org/10.1145/3526113.3545702

[28] Kent Lyons, Christopher Skeels, Thad Starner, Cornelis M Snoeck, Benjamin A Wong, and Daniel Ashbrook. 2004. Augmenting Conversations Using Dual-Purpose Speech. In *Proceedings of the 17th Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, NY, USA, 237–246. https://doi.org/10.1145/1029632.1029674

[29] Richard E Mayer. 2002. Multimedia Learning. In *Psychology of Learning and Motivation*. Vol. 41. Elsevier, New York, NY, USA, 85–139.

[30] Harry McGurk and John MacDonald. 1976. Hearing Lips and Seeing Voices. *Nature* 264, 5588 (1976), 746–748.

[31] Christian Meurisch, Cristina A Mihale-Wilson, Adrian Hawlitschek, Florian Giger, Florian Müller, Oliver Hinz, and Max Mühlhäuser. 2020. Exploring User Expectations of Proactive AI Systems. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–22. https://doi.org/10.1145/3432193

[32] Rada Mihalcea and Chee Wee Leong. 2008. Toward Communicating Simple Sentences Using Pictorial Representations. *Machine Translation* 22, 3 (2008), 153–173.
[33] Florian Müller, Sebastian Günther, Azita Hosseini Nejad, Niloofar Dezfuli, Mohammadreza Khalilbeigi, and Max Mühlhäuser. 2017. Cloudbits: Supporting Conversations through Augmented Zero-Query Search Visualization. In *Proceedings of the 5th Symposium on Spatial User Interaction* (Brighton, United Kingdom) *(SUI '17)*. Association for Computing Machinery, New York, NY, USA, 30–38. https://doi.org/10.1145/3131277.3132173
[34] Hiromu Ogawa and Pattie Maes. 2020. Smartwatch-Based Topic Suggestions to Enrich Casual Conversations in Awkward Encounters. In *Proceedings of the 2020 International Symposium on Wearable Computers*. Association for Computing Machinery, New York, NY, USA, 68–72. https://doi.org/10.1145/3410531.3414310
[35] Yi-Hao Peng, Ming-Wei Hsi, Paul Taele, Ting-Yu Lin, Po-En Lai, Leon Hsu, Tzu-chuan Chen, Te-Yen Wu, Yu-An Chen, Hsien-Hui Tang, and Mike Y. Chen. 2018. SpeechBubbles: Enhancing Captioning Experiences for Deaf and Hard-of-Hearing People in Group Conversations. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3173574.3173867
[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, Association for Computing Machinery, New York, NY, USA, 8748–8763.
[37] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *ArXiv* abs/2204.06125 (2022), 10.
[38] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *CoRR* abs/1908.10084 (2019), 10. arXiv:1908.10084 http://arxiv.org/abs/1908.10084
[39] Bradley Rhodes and Thad Starner. 1996. Remembrance Agent: a Continuously Running Automated Information Retrieval System. In *The Proceedings of the First International Conference on the Practical Application of Intelligent Agents and Multi Agent Technology*, Vol. 1. Association for Computing Machinery, New York, NY, USA, 487–495.
[40] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. https://doi.org/10.48550/ARXIV.2205.11487
[41] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. 2021. Image super-resolution via iterative refinement. *arXiv preprint arXiv:2104.07636* 1 (2021), 10.
[42] Nazmus Saquib, Rubaiat Habib Kazi, Li-Yi Wei, and Wilmot Li. 2019. Interactive Body-Driven Graphics for Augmented Video Performance. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300852
[43] Yang Song and Stefano Ermon. 2020. Improved techniques for training score-based generative models. *Advances in neural information processing systems* 33 (2020), 12438–12448.
[44] Salvador Soto-Faraco, Alan Kingstone, and Charles Spence. 2003. Multisensory Contributions to the Perception of Motion. *Neuropsychologia* 41, 13 (2003), 1847–1862.
[45] Barry E Stein and M Alex Meredith. 1993. *The Merging of the Senses*. The MIT press, New York, NY, USA.
[46] William H Sumby and Irwin Pollack. 1954. Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America* 26, 2 (1954), 212–215. https://doi.org/10.1121/1.1907309
[47] S Shyam Sundar. 2000. Multimedia Effects on Processing and Perception of Online News: a Study of Picture, Audio, and Video Downloads. *Journalism & Mass Communication Quarterly* 77, 3 (2000), 480–499.
[48] Hanna M. Wallach. 2006. Topic Modeling: Beyond Bag-of-Words. In *Proceedings of the 23rd International Conference on Machine Learning* (Pittsburgh, Pennsylvania, USA) *(ICML '06)*. Association for Computing Machinery, New York, NY, USA, 977–984. https://doi.org/10.1145/1143844.1143967
[49] Morton Wiener and Albert Mehrabian. 1968. *Language Within Language: Immediacy, a Channel in Verbal Communication*. Ardent Media, New York, NY, USA.
[50] Haijun Xia. 2020. Crosspower: Bridging graphics and linguistics. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, NY, USA, 722–734.
[51] Haijun Xia, Jennifer Jacobs, and Maneesh Agrawala. 2020. Crosscast: Adding Visuals to Audio Travel Podcasts. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 735–746. https://doi.org/10.1145/3379337.3415882
[52] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. 2022. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. https://doi.org/10.48550/ARXIV.2206.10789
[53] Jezia Zakraoui, Moutaz Saleh, and Jihad Al Ja'am. 2019. Text-to-picture tools, systems, and approaches: a survey. *Multimedia Tools and Applications* 78 (2019), 22833–22859.
[54] Xiaojin Zhu, Andrew B Goldberg, Mohamed Eldawy, Charles R Dyer, and Bradley Strock. 2007. A Text-to-Picture Synthesis System for Augmenting Communication. In *AAAI*, Vol. 7. Association for Computing Machinery, New York, NY, USA, 1590–1595. https://doi.org/10.5555/1619797.1619900
[55] Douglas E. Zongker and David H. Salesin. 2003. On Creating Animated Presentations. In *Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (San Diego, California) *(SCA '03)*. Eurographics Association, Goslar, DEU, 298–308. https://doi.org/10.1109/3DUI.2013.6550198
[56] Changqing Zou, Haoran Mo, Chengying Gao, Ruofei Du, and Hongbo Fu. 2019. Language-Based Colorization of Scene Sketches. *ACM Transactions on Graphics* 38, 6 (Dec. 2019), 233:1–233:16. https://doi.org/10.1145/3355089.3356561

A VISUAL CAPTIONS SYSTEM

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

Given the growing trend of remote work, 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 [16, 20], 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. The off-screen canvas enables prototyping of interfaces with text, image, and even 3D graphics in real time, which can be shared with other participants in video conferences, even if they don't have ARChat installed. 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 deployment to remote study participants who can enable the plugin in Chrome to gather insights from the use of Visual Captions in everyday meetings.

A.2 Settings Page

Figure 14 shows the settings page of Visual Captions.

B CROWDSOURCING TASK

B.1 Crowdsourcing Task Instruction

Figure 13 shows our task instruction and examples provided to crowdworkers before they proceed to tasks.

B.2 VC1.5K Topic Distribution

Figure 15 shows topic counts for conversation categories in the VC1.5K dataset.

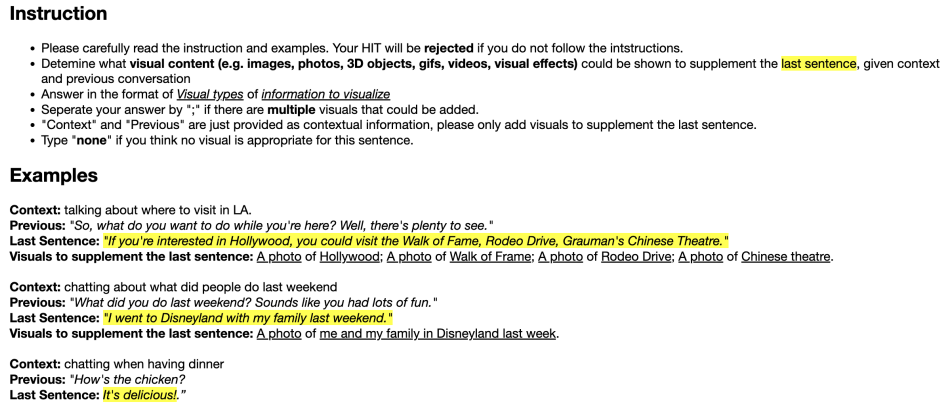

Figure 13: Task instructions and examples we provided to crowdworkers on Amazon MTurk.

C STUDY DESIGN

C.1 Technical Evaluation Survey Questions

We asked crowdworkers to rate their agreement with the following statements for each visual suggestion:

- Q1 I would like to display the visual when having the conversation.
- Q2 The displayed visual provides useful information to the conversation.
- Q3 The displayed visual is relevant to the conversation.
- Q4 The displayed visual is appealing and of high-quality.
- Q5 The displayed visual is presented in an appropriate visual type (*e.g.* image vs. emoji).
- Q6 The displayed visual is selected from a correct source (*e.g.* online search vs. personal).

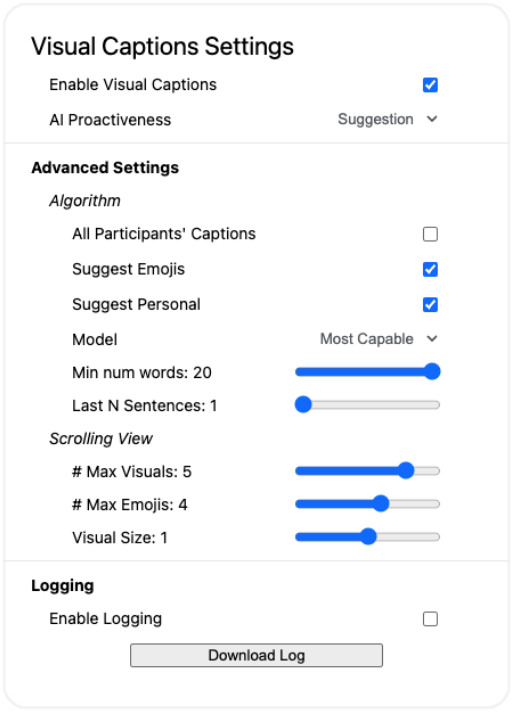

Figure 14: Visual Captions allows users to customize settings including levels of AI proactivity, whether to suggest emoji or personal images, punctuality of visual suggestions, visual suggestion models, etc.

C.2 User Study Questions

After each study condition, we ask participants the following questions:

C.2.1 *Task Load Index.*

- Mental demand: how mentally demanding was the conversation?
- Physical demand: how physically demanding was the conversation?
- Frustration: How stressed and annoyed were you during the conversation?
- (only with VC) Temporal demand: how hurried or rushed was the pace of using Visual Captions?
- (only with VC) Performance: How successful were you in selecting and displaying the desired visuals?

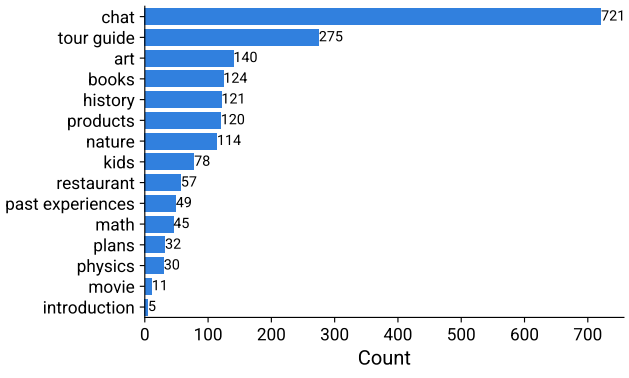

Figure 15: Topic counts for each category in the VC1.5K dataset.

- (only with VC) Effort: How hard did you have to work to select and display the visuals while having the conversation?

C.2.2 Interaction with AI.

- I feel like I have control over what image to be displayed.
- I feel such a system could infringe my privacy and is intrusive.
- The system could make a mistake that would be hard for me to override.
- The system interrupted my conversation experience.

After the open-ended conversation, we asked participants to rate their agreement with the following statements:

C.2.3 How Visuals Help?

- Knowledge: The suggested visuals can help me understand some unfamiliar terms and concepts, or make the conversation more informative.
- Clarification: The suggested visuals can help clarify ambiguity in the conversation.
- Engagement: The suggested visuals can make the conversation more interesting and enjoyable.
- Creativity: The suggested visuals provided topics to talk about or led the direction of the conversation.

D USER STUDY RESULTS

We visualize the detailed distribution of participants' answers to Task Load Index and Likert-scale questions, comparing the usage of Visual Captions and different modes of Visual Captions.

	Mental Demand	Physical Demand	Frustration	Temporal Demand	Effort	Performance	Control	Intrusive	Error-prone	Interruptive
No VC	2.1 (1.34)	1.7 (1.19)	2.1 (1.45)	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Auto-Display	2.25 (1.44)	1.7 (1.23)	2.45 (1.6)	2.65 (1.35)	1.35 (0.73)	3.85 (1.31)	2.55 (1.32)	3.85 (1.65)	5.65 (1.24)	4.15 (1.56)
Auto-Suggest	3.25 (1.22)	2.85 (1.77)	2.05 (1.2)	3.7 (1.45)	4.05 (1.07)	4.8 (1.4)	4.9 (1.67)	2.5 (1.6)	2.7 (1.62)	3.65 (1.28)
On-demand-Suggest	3.95 (1.56)	4.05 (1.8)	3.45 (1.99)	2.55 (1.12)	4.4 (1.28)	3.85 (1.35)	5.15 (1.28)	2.35 (1.28)	2.6 (1.28)	4.75 (1.37)

Table 1: Means (standard deviations) of participants ratings to user study questions.

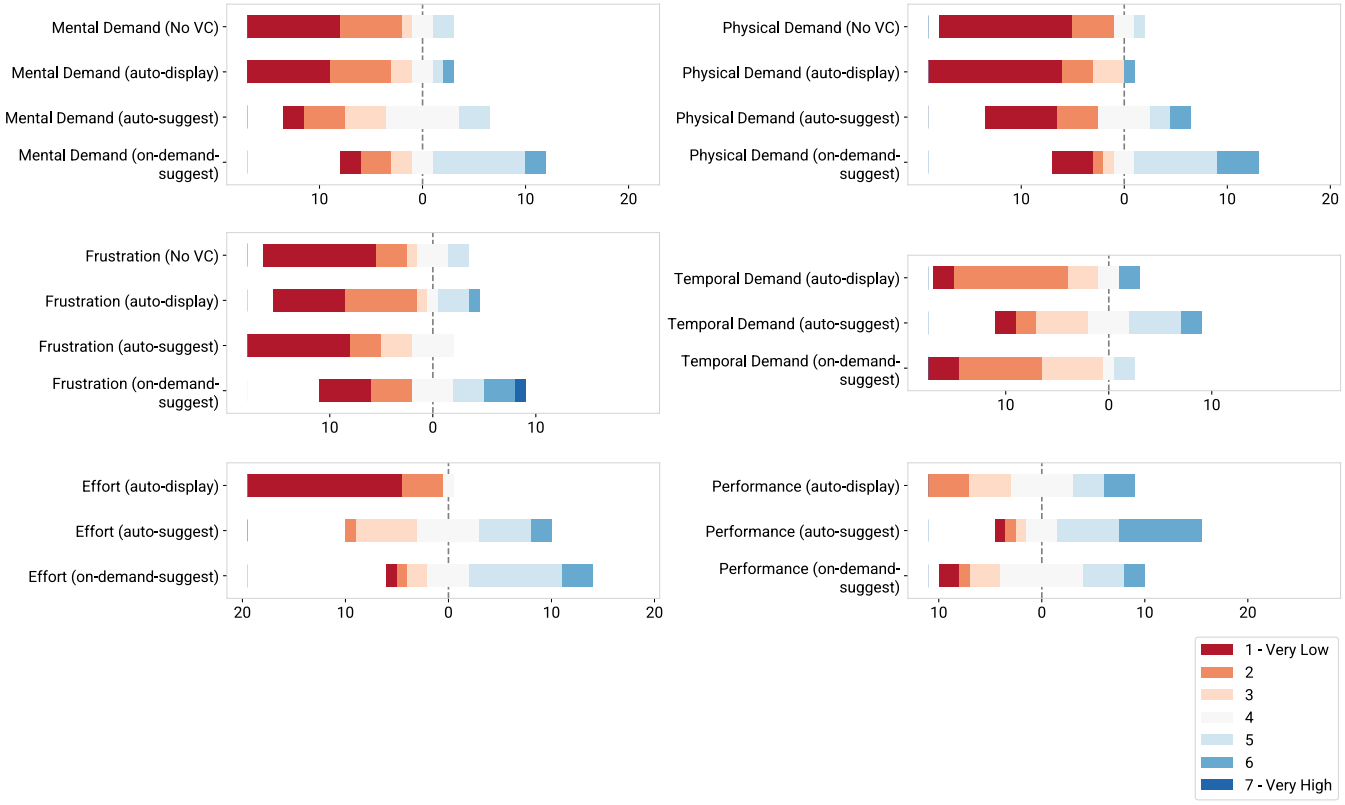

Figure 16: Detailed distribution of participants' answers to Task Load Index questions.

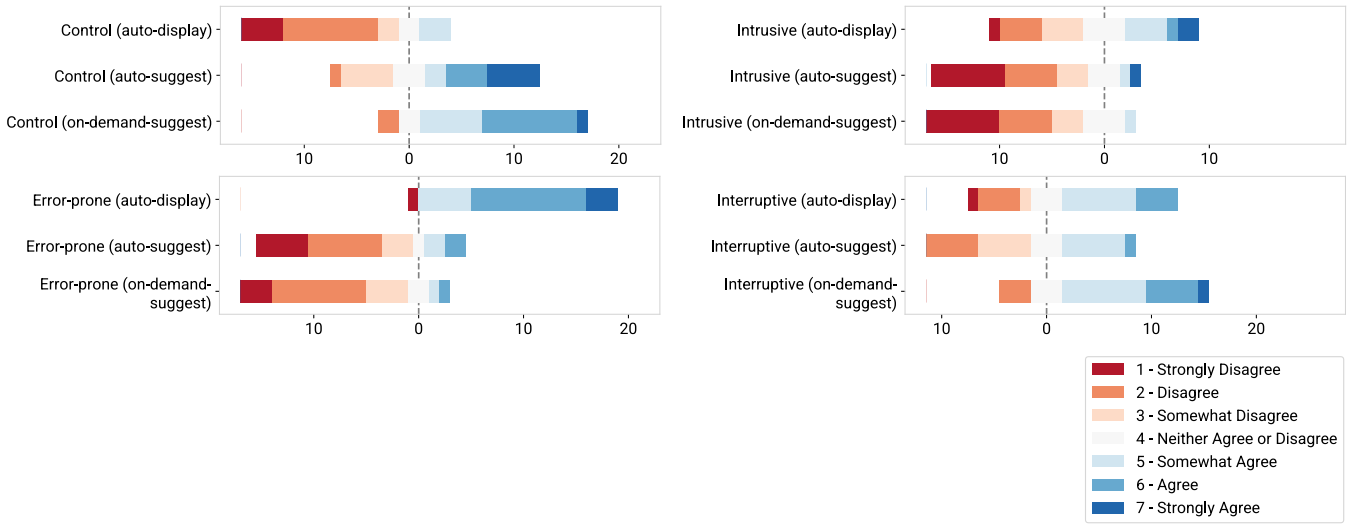

Figure 17: Detailed distribution of participants' answers to Likert-scale questions, comparing different modes of Visual Captions.