Augmenting Conversations With Comic-Style Word Balloons

Heng Zhang, Lifeng Zhu, *Associate Member, IEEE*, Qingdi Chen, Aiguo Song, *Senior Member, IEEE*, and Lap-Fai Yu

***Abstract*—We propose a novel approach for enabling comic-style conversation in mixed reality to assist face-to-face conversation on-site or remotely. Our approach brings word balloons of comic-style conversation to the real world. The word balloons can adapt to mixed reality scenes, such as the 3-D head motion of the speaker, the comic styles, and the speech. During the conversation, our approach updates the word balloons continuously in the object space and discretely in the image space, guided by a field learned from comics. Quantitative experiments and perceptual studies were conducted to evaluate and compare our approach with alternatives. The results from the user study and ablation study demonstrated that our approach turns out to be practical for assisting face-to-face conversation.**

***Index Terms*—Comics, mixed reality, word balloons.**

I. Introduction

FACE-TO-FACE conversation is one of the most dominant ways of interpersonal communication. Although tools such as instant messaging and video conferencing are easily accessible and frequently used daily, the rich nonverbal information implicitly displayed in a face-to-face chat makes it functionally irreplaceable by popular remote conversation approaches. To avoid misunderstanding, attention should be paid to carefully expressing, hearing, understanding, and recalling the speech. It is promising to develop a novel tool to assist conversation by improving the quality and efficiency of interpersonal communication.

Comics are a popular and appealing medium to present an imaginary world. In comics, graphic novels in combination with texts efficiently present scripts. The compelling storytelling nature makes comics applicable as a medium for presenting information or as a tool for general interpersonal communication. In comic worlds, conversations are presented as texts encapsulated in a blank region overlaid on the panel, called word balloons or speech bubbles. As one of the most distinctive elements of the comic medium, a word balloon intuitively associates the words with the speaker and conveys the contents of the dialogue [1]. In comics, text and visual artifacts compensate each other to create a compelling and powerful storytelling medium [2]. Inspired by the comic medium, we propose to learn from the word balloons in comics and use them to enhance face-to-face conversation and video chat.

In transferring the word balloons in comics to the real world in mixed reality (MR), we mainly consider the size and placement of the word balloons. Other attributes, such as color, font, or decoration, could be prespecified. At the same time, the placement and size cannot be prespecified because the camera and the scene vary continuously in the real world. The related geometry of the word balloon should be dynamically updated. Previous studies on comics have demonstrated the possibility of automatically creating static word balloons or the layout with an offline process. Some virtual reality applications also have pre-programmed dialogues based on acquired knowledge of the characters' positions in advance. In an MR scenario, characters and other objects in a real-world environment may move dynamically with the view update, which makes it significantly different from static or other systems that assume the scripts or even the characters' motions are known beforehand. We want the word balloons to react to the actual environment, follow people's movements, and not hinder the visual clues in the environment in real time, which requires careful consideration and design. Technically, there are two main challenges to address: 1) to determine where to put the word balloons; 2) to make the word balloons user-friendly and practical.

Our approach complements MR [3], [4] by combining the pure virtual environment in comics and the natural environment in the physical world. In MR, users can interact with virtual objects, i.e., word balloons. Fig. 1 shows real-world conversation augmented by our generated word balloons. e demonstrate merging the static 2-D comic world with the dynamic 3-D real world and evaluate our approach via a set of user studies. We investigate if it is natural to put word balloons in the real world, whether the word balloons improve interpersonal communication and the differences between word balloons and subtitles. Our main technical contribution is a pipeline that combines a learning-based guidance field and a force-based physical model to dynamically layout the word balloons in response to the open MR world.

Manuscript received 1 March 2022; revised 26 June 2022, 24 September 2022, and 14 November 2022; accepted 19 November 2022. Date of publication 2 December 2022; date of current version 15 March 2023. This work was supported in part by the Basic Research Project of Leading Technology of Jiangsu Province under Grant BK20192004, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20211159, and in part by the Fundamental Research Funds for the Central Universities. This article was recommended by Associate Editor G. Serra. *(Corresponding author: Lifeng Zhu.)*

Heng Zhang, Lifeng Zhu, Qingdi Chen, and Aiguo Song are with the State Key Laboratory of Bioelectronics, Jiangsu Key Lab of Remote Measurement and Control, School of Instrument Science and Engineering, Southeast University, Nanjing 211189, China (e-mail: mediosrity@gmail.com; lfzhulf@gmail.com; chin20191@fuji.waseda.jp; a.g.song@seu.edu.cn).

Lap-Fai Yu is with the Department of Computer Science, George Mason University, Fairfax, VA 22030 USA (e-mail: craigyu@gmu.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/THMS.2022.3224767.

Digital Object Identifier 10.1109/THMS.2022.3224767

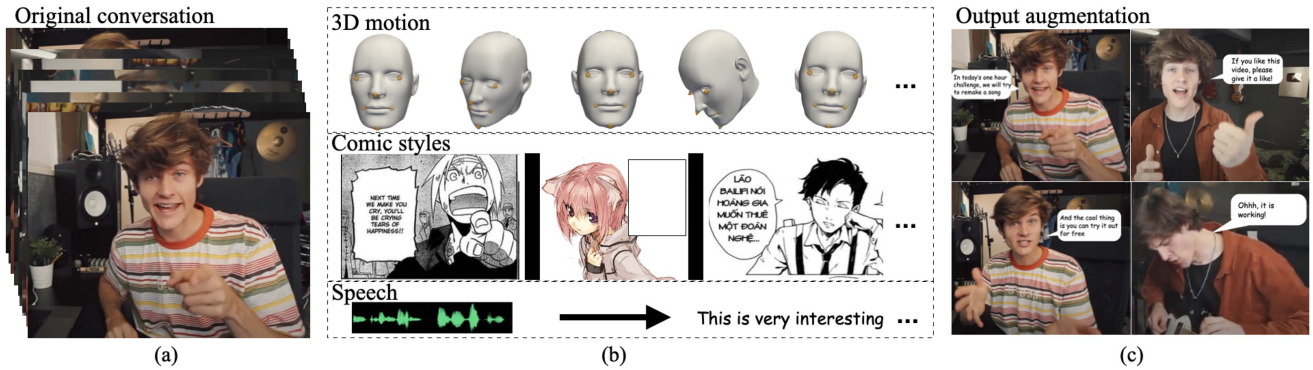


Fig. 1. Our approach automatically augments real-world conversation with comic-style word balloons. (a) Original conversation in the real world. (b) Our approach considers the 3-D head motion of the speaker, comic styles, and the speech. (c) It places comic-style word balloons at appropriate places in the scene automatically.

Specifically, our contributions include the following:

1) a learning-based guidance field which provides a prior prediction of the possible positions of word balloons;
2) a force-directed layout, which makes the word balloons not overlap with other nonverbal information like hand gestures and visual clues in the background.

II. Related Work

A. Comics and Word Balloons

Research about comics has been conducted in different research fields for years [5], [6]. This work will focus on the comic dialogues presented in word balloons. In the early work of Kurlander et al. [1], word balloons were synthesized to enhance the experience in online chat rooms. A greedy method was used to make the space for the word balloon in the comic panel. The layout of the word balloons is further studied by Chun et al. [7], where the authors first placed the word balloons near the actors and then refined the layout with a heuristic method. Toyoura et al. [8] proposed a rule to divide the panel based on eye-tracking data and position word balloons in the most significant region. Cao et al. [9] considered the reading order and synthesized manga layouts, including the word balloons, through a probabilistic graphical model.

Researchers also associated comic-style word balloons with real-world photos and videos. In the work of Chen et al. [10], word balloons were manually placed on a comic strip with an interactive sketch-based user interface. A manga-style layout was also proposed to be automatically generated from a video [11], [12]. The quality of the layout, including the placement of word balloons, was measured with an extracted salience model. Then a sampling algorithm was proposed to optimize the layout. All these works are primarily on word balloon synthesis on static images. Greedy and heuristic algorithms do not consider the dynamic transition of the word balloons.

B. Labels and Captions in MR

As we will study comic-style talk in MR, the word balloons should be well aligned with the physical world in real time. A recent work called SpeechBubbles [13] by Peng et al. has proved that the word balloons overlaid in the real world are helpful in group conversations for deaf and half-of-hearing people. Later, Kurahashi et al. [14] studied the strategy to separate the word balloons when the speeches are long. Similar work has also been proposed to put captions along the trajectory of the tracked speaker by Kushalnagar et al. [15], and by Rothe et al. [16]. In the work of Kurzhals et al. [17], another strategy to control the placement of captions was proposed to use gaze as the input, where the captions follow the user's gaze on the screen. Although word balloons in MR have been presented, the word balloons are just positioned in image space. Only a simple rising motion is applied to the balloons in the interaction. In this work, we will further exploit the immersive feature of MR and study how to align the word balloons dynamically with the 3-D world. We also aim to design innovative word balloons to avoid obstructing the scene, keeping other nonverbal information in a conversation visible to the users. In addition, SpeechBubbles [13] focused on testing the design options in placing words into the bubbles. In contrast, we focus on merging the word balloons in comics into the real world, visualizing the conversation in MR.

In MR, both the speakers and the camera may move frequently. The presentation of subtitles in virtual reality (VR) has been studied by Sidenmark et al. [18]. Proper attachment of the word balloons to the speakers is similar to the task of annotation or labeling in augmented reality (AR), which was studied by Bell et al. [19]. In the work of Orlosky et al. [20], labels are defined following specific patterns. Bell et al. [19] updated the position of the labels on a real-world object according to the view change with a greedy placement and Pick et al. [21] used a force-based approach. Strategies to guide the placement of the labels in 2-D or 3-D have also been proposed by Grasset et al. [22] and Tatzgern et al. [23], respectively. Madsen et al. [24] studied different updating schemes in different spaces and proposed to place the labels in 3-D with a discrete update.

C. Speech Processing in MR

Our work also relates to speech and sound processing in augmented or mixed reality. As a continuum between virtual and

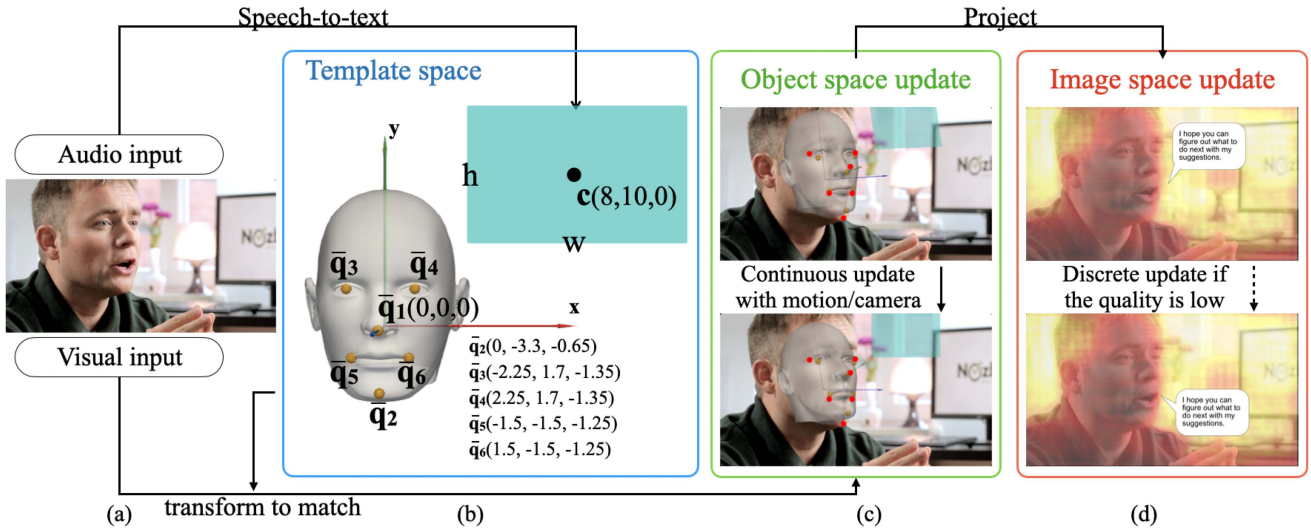


Fig. 2. Pipeline of our approach. (a) Input to our approach. (b) Unknown variables $\mathbf{c}$, w, h and their initialization in template space. The audio input is converted to texts and mapped to the template space. Then the predefined landmarks $\bar{\mathbf{q}}_i$ in the template space are aligned to the visual input. (c) Aligned word balloons are continuously updated in the 3-D object space as the head or the camera moves. (d) If the quality of the word balloon projected in the image space is low, our approach triggers image space update discretely to refine the word balloon placement.

real environments [3], MR enables interactions more than those in visual channel [4]. In the work of Chen et al. [25], the sound and speech were processed and set as the input commands in MR. Speech processing inspired and guided by comics is also taken to AR recently. Wang et al. [26] processed the background sound in the video and visualized the video sounds with comic-like texts. The shape of the texts was animated with the video to enrich the user experience.

III. Approach

A. Overview

1) Design Rationale: Our work aims to introduce comic-style word balloons to facilitate daily conversions with MR. As pointed out by [27], visual languages in VR/AR broaden the written language and facilitate interpersonal communication. It is also reported that people learn better when on-screen words are placed next to the corresponding part of the graphic (spatial contiguity) and when corresponding narration and graphics are presented simultaneously (temporal contiguity) [28].

Compared to subtitles, word balloons commonly found in comics excel at delivering information. Inspired by such findings, we designed an MR interface to augment conversation using comic-style word balloons. To create a natural and comfortable user experience, we place word balloons at appropriate locations next to the speaker by learning from comics.

To improve the quality and efficiency of conversation, the design of the word balloons follows these guidelines: 1) the word balloon acts as a balloon in the real world attached to the speaker so that it follows the speaker's motion well in the 3-D world; 2) the word balloon stably faces the camera and updates its scale with the detected speech so that it well presents the words during the conversation; 3) the word balloon does not overlap the facial emotions, gestures, or other important elements in the background so that it does not obstruct other nonverbal information in the conversation.

Aligning the word balloons in comics to the real world is still challenging because the data are from two different sources. First, word balloons in comics are displayed on 2-D images, while the scenes in the real world are 3-D in nature and have complex structures. Second, words in comics are known in prior and displayed on static images, while the conversation in the real world is dynamically changing. It is unclear how to adapt the imaginary data to the real world or make it helpful in our daily conversation. It is nontrivial to design a comic-style conversation in MR.

In this work, we involve both the design considerations in MR and comics to prototype our approach. The pipeline of our approach is shown in Fig. 2. In our design, we distinguish two spaces, object space, and image space. The object space includes 3-D landmarks and enables the 3-D transformation of word balloons before they are projected onto the screen. On the other hand, the image space has a 2-D layout of the word balloons overlaid on the 2-D pixels and only allows translation of the balloons. SpeechBubbles [13] only models the word balloons in the image space and may have a drifting artifact if the speakers are moving. As suggested in the work of Madsen et al. [24], the view management of labels in the object space outperforms those in the image space. However, it is expensive to reconstruct everything in the conversation captured by the cameras into the 3-D object space. Therefore, we consider both the object and image space in our design.

Specifically, we first work in the object space by considering the features in MR and make the word balloons stably attach to the speaker (Section III-B). Then the word balloons will be updated in the image space by considering the features learned from comics and balance with other visual information (Section III-C). We will also detect the speech and model it

into our approach with other features designed for interpersonal conversation (Section III-E).

B. MR-Style Balloon Placement

To make the word balloon well align with the speaker, we first place them in the object space so that its motion rigidly follows the speaker in the 3-D space. In this step, we introduce a template space and parameterize the unknown variables in the template space. As illustrated in Fig. 2(a), we set nose, chin, left corner of left eye, right corner of right eye, left corner of mouth, and right corner of mouth as the six landmarks on a template face. By aligning the template face with the *xy*-plane of the template space, we set the default 3-D coordinates $\bar{\mathbf{q}}_i (i = 1, 2, \ldots, 6)$ of the landmarks by the measurements on an average human face [29]. The template face model will be transformed to match the detected face in the real world. By setting default camera projection operation P, we use the PnP model [30] to find the optimal rotation R and translation $\mathbf{t}$ of the template face by minimizing $\sum_{i=1}^{6} ||P(R\bar{\mathbf{q}}_i + \mathbf{t}) - \mathbf{q}_i||$, where $\mathbf{q}_i$ is the detected landmarks of the faces on the image. We use the model in the work [31] to track the face and the landmarks on it. In this case, we match the 3-D template face with the detected face in the real world. Note that other registration method can be used as well.

We are now ready to locate the word balloons in the object space. Although the output of our approach is the location of each balloon in the image space $\{x_i, y_i\}$, we do not directly update them because the word balloons need to follow the 3-D motion of the speaker. We instead work on the template space and parameterize the location of each word balloon in the template space with its center $\bar{\mathbf{c}}_i = \{\bar{x}_i, \bar{y}_i, 0\}$ and size $\{\bar{w}_i, \bar{h}_i\}$. By default, we set the location of the word balloon to the upper right of the nose as $\bar{x}_i = 8\,cm, \bar{y}_i = 10\,cm$ and its width $\bar{w}_i = 10\,cm$. Its height $\bar{h}_i = n \cdot h_0$, where n is number of the lines of the texts and $h_0 = 2\,cm$ matches the default size of the word in the balloon, which will be discussed in Section III-E.

After setting the word balloon as a rectangle $\bar{B}$ in the template space, we apply the rotation R and translation $\mathbf{t}$ to each word balloon and get its location in the image space $P(R\bar{\mathbf{c}}_i + \mathbf{t})$. In this case, each word balloon will rigidly follow the speaker in the object space without any drifting problems. We then rotate the word balloon to make it face the camera and compensate for its tilt as a postprocess. Specifically, we first record the aspect ratio of every balloon. For every frame, we rotate each balloon to create a new virtual one and get its scale factor. Then we rescale the size of a balloon to fit the balloon which correctly faces the camera, keeping the aspect ratio.

After our treatment in the object space, the word balloon acts like a real balloon attached to the speaker in a MR-style. It aligns with the speaker and moving camera, as shown in Fig. 2(c). We also set more candidate positions $\bar{\mathbf{c}}_i = \{\pm\bar{x}_i, \pm\bar{y}_i\}$. If the default position does not well fit the screen space, we use another candidate as the initialized position. As the speaker may frequently move when speaking, the word balloons may move too much due to head pose estimation noises, making them difficult to read. To stabilize the movement of the word balloons, we propose to update the word balloon placement discretely rather than continuously. Specifically, when a word balloon is displayed, we record its corresponding head pose $(R, \mathbf{t})$. We frequently check the difference between the new pose $(R,' \mathbf{t}')$ and the original head pose $(R, \mathbf{t})$, and only update the word balloon placement when the difference exceeds a threshold. In our implementation, the thresholds of the rotational angle and translation are set as 60° and 20 cm, respectively. If the face is occluded, we do not update the word balloon placement. We simply put the word balloon at the center of the view if no face is detected in the initialization.

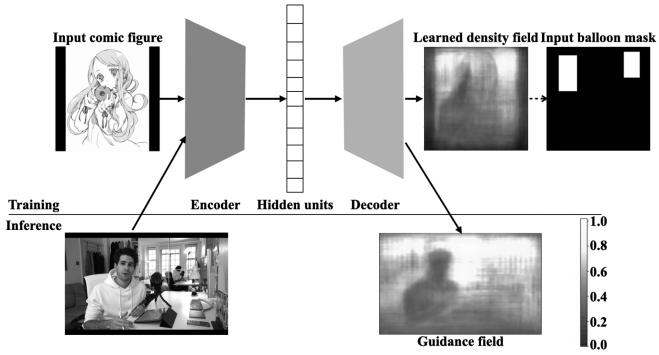


Fig. 3. We take a set of comic figures and the masks extracted from the corresponding annotated word balloons as the input to train a neural network. The network outputs a learned density field to indicate the probability of the word balloon placement, which is called a guidance field. By feeding a photo of a real-world conversation into the network, our approach obtains a guidance field for placing word balloons. The color bar indicates the probability of the word balloon placement.

C. Comic-Style Balloon Placement

Although the word balloons obtained in the first step satisfy the requirements of MR, the balloons may overlap or obstruct crucial visual information. In this step, we will make the word balloons intelligent so that the visual information in the conversation is not hindered. However, it is an open problem to reconstruct motion and detect salience from captured image sequences in real time. Because the word balloons are virtual objects we bring from comics, we propose to learn the placement of the balloons from comics. Our insight is that the word balloons in comics are aware of the comic layout, so they do not obstruct the essential visual information. Besides, the word balloons in comics are also aware of the speakers. The features we learned from comics provide a good balance between the constraints that the balloons are near the speaker and that the balloons do not obstruct crucial visual information. We first introduce a guidance field that we learned from comics. Then we present our method to refine the location of word balloons.

D. Learning Guidance Fields From Comics

Finding possible regions in photos to place a word balloon is similar to a region proposal problem [32]. Here, we learn the relationship between word balloons and the scenes in comics and use it to create a guidance field, as shown in Fig. 3. We design

Fig. 4. Guidance fields computed from comics and real-world photos.

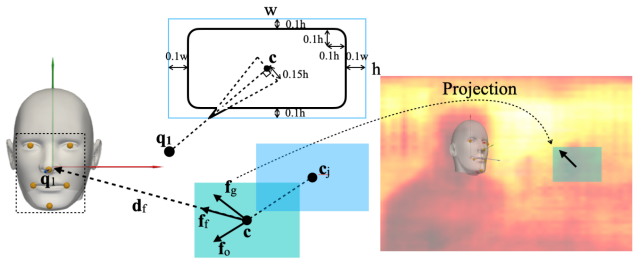


Fig. 5. Force model in the image space update. The shape of the word balloon is defined by its position and size, as well as the position of the nose landmark.

images as the input of this step. As the output of our model, we expect a scalar field representing the spatial distribution of the probability of the word balloon's location, as shown in Fig. 4.

1) Dataset: We collected 5000 manga images from Manga109 [33], which is composed of 109 manga volumes drawn by professional manga artists in Japan. The positions of texts in each panel are manually annotated. Since images in Manga109 have word balloons painted on the panels, we utilize a manga inpainting method [34] to obtain the truth background by removing dialogue balloons. However, we do not want the learning process to entirely depend on the assumptions of the image inpainting methods. We also use an existing dataset of different manga images from different comic artists [35]. In total, we select 5000 panels and each panel contains one to five comic character faces. Therefore, we take clean manga images without any word balloons on them from the Danbooru2019 dataset [35]. Although the images in the dataset are manga-style illustrations but not samples from true manga books, we make the word balloon design on them to have the same style of comics in the data preparation stage. We employ cartoonists and manga readers to annotate the possible area of the word balloons in the manga images. Specifically, they draw one or multiple rectangles on the manga images as if the cartoonists are positioning word balloons on the pages. We convert the annotated image to a mask with the pixel inside the balloons to be 1 and other pixels are set to 0. All images are preprocessed to have a resolution of 512×512 and have three color channels.

2) Learning: We use the DeepLab network [36] to train our model with the same network structure in [37]. The size of our input is $512 \times 512 \times 3$, and the output is a scalar map called the guidance field G with size $512 \times 512 \times 1$. We divided the dataset into a training set, a validation set, and a test set in the ratio of 8 to 1 to 1. The network is trained using the Momentum optimizer with a batch size of 4 samples and a momentum of 0.9. The weight decay is 0.00004, and the learning rate is 0.001 in our setting. We use cross-entropy as the loss function.

3) Evaluation: We conduct experiments to test the performance of our model. We feed 300 pages from comics and 300 real-world photos to our network and generate the corresponding guidance fields. The real-world photos are taken from the videos recorded from chatting scenes. Then we invite 24 volunteers (including five designers) to rate if they are satisfied with the guidance field representing the probability map for the word balloons. We randomly assign the image and guidance field pair to the volunteers, and Likert scores (1: poor, 5: perfect) are collected. The average scores of the comic pages and real-world photos are $M = 3.9$, $SD = 0.88$ and $M = 3.82$, $SD = 1.13$, respectively. We list the results of four input pages and photos in Fig. 4. In this work, we choose to use DeepLab because it is compatible with our mobile MR framework ARKit and its quality and performance are sufficient for our application.

4) Update With Guidance Field: After obtaining the guidance field, we design a framework to position and animate the word balloons in the image space. Following the rules proposed by Madsen et al. [24], we propose to use discrete updates. To facilitate the reading of the word balloons, we only trigger the update of the word balloons in the image space discretely. For example, they overlap or have a shallow probability score on the guidance field. If the position of the word balloon should update, we update it directly in the object space and then project it into the image space so that the word balloon still follows the motion of the speaker or camera in the MR. We restrict the balloons to move on the plane $z = 0$ in the template space and use a physically based model similar to boid [38] to make the word balloons intelligent. As illustrated in Fig. 5, we model our design considerations into forces to animate each word balloon

$$\mathbf{f} = \mathbf{f}_g + \mathbf{f}_f + \mathbf{f}_o.$$

The first force $\mathbf{f}_g$ makes the word balloons aware of the guidance field. We integrate the guidance field inside the rectangles in the image space to measure the quality of the word balloons. The force $\mathbf{f}_g$ is set as the gradient of the quality so that it will increase the probability of the word balloon according to the guidance field. Suppose the word balloon $\bar{B}(x, y)$ at (x, y) is projected to a rectangle $B(x, y)$ in the image space, we define its quality under the guidance field as $e(x, y) = \sum_{(i,j)\in B(x,y)} G(i, j)$, where $G(i, j)$ is the value of the guidance field at (i, j). Then we use the finite-difference to approximate the gradient of $e(x, y)$ and model the first force

$$\mathbf{f}_g(x, y) = \omega_g \left(\frac{e(x + \delta, y) - e(x - \delta, y)}{2\delta} \right.$$
$$\left. \frac{e(x, y + \delta) - e(x, y - \delta)}{2\delta} \right)$$

where ω_g is the weight. We only evaluate and apply this force when the quality $e(x, y)$ is smaller than a threshold e_0 according to our discrete update rule.

The second force $\mathbf{f}_f$ penalizes the distance from the word balloon to the speaker if the balloon overlaps with the speaker's

face or is too far away from the speaker. We use the landmark of the nose $\bar{\mathbf{q}}_1$ and a rectangle $\bar{B}_f$ that encloses the face model in the template space to represent the face. Let $\mathbf{d}_f$ denote the direction from the center of the word balloon to $\bar{\mathbf{q}}_1$, we set $\mathbf{f}_f = \omega_f \mathbf{d}_f$ if the distance between the word balloon and the speaker is greater than a threshold l or $\mathbf{f}_f = -\omega_f \mathbf{d}_f$ if the word balloon $\bar{B}$ overlaps with the face $\bar{B}_f$, where ω_f is the weight.

The last force acts on the word balloons if they overlap with each other. Suppose the ith word balloon $\bar{B}_i$ overlaps with the jth word balloon $\bar{B}_j$, we set $\mathbf{f}_{oi} = -\omega_o a_{ij} \mathbf{d}_{ij}$ and $\mathbf{f}_{oj} = \omega_o a_{ij} \mathbf{d}_{ij}$, where ω_o is the weight and $\mathbf{d}_{ij}$ is the direction from the center of $\bar{B}_i$ to that of $\bar{B}_j$ and a_{ij} is the overlapping area $|\bar{B}_i \cap \bar{B}_j|$.

We sum up all the forces on the word balloon if a discrete update is triggered. Then we integrate the force to update the balloon's velocity and obtain its new position by integrating the velocity. After each update step, we damp the velocity with a coefficient λ to make the animation stable. If the word balloon is out of the image space, we apply an extra force to push it back to the image space. After we update the word balloon in the template space, we project the rectangles to the image space and render the word balloon according to the projected rectangle with a simple rule, as illustrated in Fig. 5. Note that we do not apply hard constraints to avoid occlusions. After a smooth animation, the proposed method will respond to the occlusions and animate the word balloons for better placement.

E. Speech-Aware Design for Conversation

We also design the word balloons to be aware of the speech in the conversation. We simultaneously run speech processing and update the word balloons and the visual information in the previous subsections. We test the iFLYTEK [39] and Google Cloud Speech-to-Text API [40] to convert the speech into texts. Other real-time speech-to-text tools such as [41] may also be applied to our system. The speech-to-text module outputs the detected texts in real time. Based on the outputs from the speech recognition module, we update the word balloons in the template space. In this way, the speaker's motion, and the view update are decoupled from the audio channel. This speech-aware design does not conflict with the MR-style and comic-style balloons. Note that the speech-to-text module does not predict the transcription. There is also a lag between text and speech because speech-to-text recognition cannot be performed before the speech completes. In our design, we remind the users of this latency with an updating ellipsis at the end of the texts to imply the ongoing transcription status.

F. Drawing Word Balloons

The word balloons are represented as rectangles in the previous sections. After we place each word balloon according to the rectangle, we define its shape and draw them in the image space. In our implementation, the rules for mapping the rectangle to the balloon shape are as illustrated in Fig. 5. Another modification of the rules is also allowed in our framework. We fill the texts inside the word balloon in the template space with the default size (the height of one character $h_0 = 2$ in our implementation). The actual size of the words in the image space scales with the word balloons, similar to the textures on the word balloons. In this way, if the speaker is moving far away, the words in the balloon are getting smaller, following the nonverbal information in the conversation and what we usually observe in comics. We modify the balloon style to highlight the newly generated one to the user.

TABLE I
DEFAULT SETTINGS OF CONTROLLABLE PARAMETERS

Parameter	e_0	ω_g	ω_f	ω_o	l/cm	λ	fs	t_v
Default	0.3	1.3	25	30	40	0.5	16	5 s

TABLE II
AVERAGE COMPUTATIONAL COST

Module	Inference	Speech2Text	Position update
Time	43 ms	200 ms per word	3 ms

G. Multiple Speakers

If there are multiple speakers in the conversation, speaker recognition [42] can be used to label the speakers from audio inputs. Because the speaker diarization feature from commercial speech-to-text tools only returns the results in the final response, they are not mature enough to support real-time streaming data. Therefore, we restrict dealing with cases when multiple speakers take turns speaking. In this case, we use the break between the voices of different speakers to separate the speech. After getting the recognition results separately, we generate word balloons for every speaker with the words they said. To map the word balloon to the right speaker, we use the method in [43] to find the speaker in the scene. Furthermore, if the voices are from the audience, we skip the speech-to-text operation on them.

IV. IMPLEMENTATION

We implement our approach with the Apple ARKit and test it on the iPad Pro 11 inch. We tune the parameters in our method using the prototype system and find the default parameter settings, as listed in Table I. e_0 is the threshold to trigger image space update; ω_g, ω_f, and ω_o are the weights in computing the force; l denotes the distance threshold to trigger the force $\mathbf{f}_f$; λ is the damping coefficient of the velocity; fs is the font size, and t_v denotes the life span of a word balloon before it fades out in seconds. We time each component of our approach, as shown in Table II, including the time for computing the guidance field (inference), converting speech to words (Speech2Text), and updating the position of each word balloon (position update). The components for computing the guidance field, converting speech to text, and updating the word balloons run at different threads in our approach. The computation of the guidance field is relatively expensive. However, because the continuous update in view management does not perform better than the discrete update [24], we do not compute the guidance field every frame. In our implementation, we refresh the scene at a rate of 60 frames per second and update the guidance field every 30 frames. The image space update also does not operate each frame. It runs when the balloons are at the low probability region in the

Fig. 6. Word balloon naturally attached to the speaker without any drifting artifacts with the object space update. When the speaker moved closer to the camera, the word balloon became bigger.

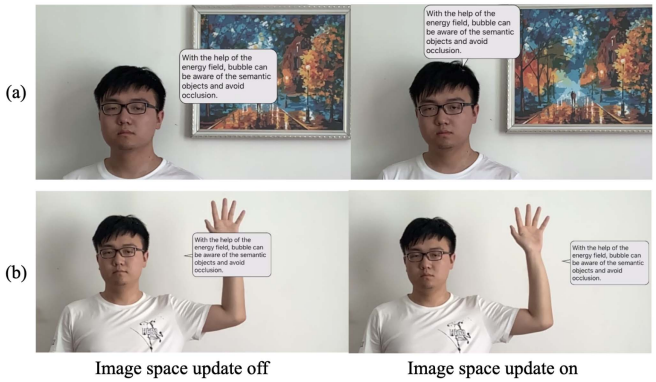


Fig. 7. With the image space update, the word balloons are aware of (a) the background and (b) the gestures and will not obstacle these nonverbal information.

guidance map, overlap with each other, and are newly added or fading. We run the image space update for every 200–300 frames on average for a typical scenario. In general, the entire approach works in real time.

V. Evaluation

A. Ablation Study

In this section, we conduct an ablation study to evaluate the contribution of each component in our approach.

1) Object Space Update: In Fig. 6, we remove the object space update and show the results from only image space update. Compared with the results from the full pipeline, the word balloons are floating around the speaker but are not attached to the speaker's head anymore. The speaker stays at a distance of about 1.5 m to the camera, and when he walks or rotates his head, the balloon will not move immediately with the movement of the head but goes closer to the head under the influence of the forces. The balloon will change size when the speaker walks close or far away from the camera. When the speaker rotates his head, the balloon will rotate as well. With the object space update, the balloons are more like virtual objects rather than graphical contents on the screen.

2) Image Space Update: In Fig. 7, we disable the image space update and have the word balloons only update with the speaker's face and camera motion. Compared with the results from the entire pipeline, the word balloons are still attached to the speaker's head rigidly and act like the actual balloon. However, when the speaker acts on gestures, such as waving hands, the balloon will not move to a different position and may cause an occlusion. With the image space update, the word balloon will recognize some nonverbal information in the scene, such as the gestures made by the speaker and the decorations in the background, which may contain some semantic information.

B. User Study

1) Study Design: Our user study contains two parts, and the independent variable is augmentation mode. We designed three augmentation modes: the view with recognized texts shown as the subtitle (S), the view with recognized texts shown as floating speech bubbles only updating in image space (F), and the view with recognized texts shown with the proposed comic-style balloons in MR (M). In mode F, we annotated the word balloons only in the image space without the guidance field, which was an implementation of the speech bubble [13]. Because mode M differs from mode S and F in terms of the animated layout of the word balloons, we focused on this difference and only had one speaker in the view to cancel other factors. Because the effects of object space update had been validated [24], we did not separately test the object space update in the user study.

Participants were invited to experience the system in pairs in the first part. One participant acted as a speaker in each pair, and the other was the audience. The audience held a tablet running our application and was in front of the tablet's screen (with 2732×2048 resolution). At the same time, the audience wore a pair of earplugs inside a headset to simulate the loss of hearing. For each trial, the speaker was asked to give a 4-min introduction of himself/herself in front of the tablet's camera. The speaker was free to use his/her body language, including moving the body or rotating the head. The audience could not listen to the voice but see the action and the virtual word balloons. In each trial, we sequentially switched the three modes, S, F, and M, and each mode ran for one minute, and the raw video clip without augmentation was saved. At the last minute, the audience who held the tablet was free to switch the three modes. After the experiment, they were required to rate the three different modes. We collected subjective feedback on a 5-point Likert scale (1: strongly disagree, 5: strongly agree). Four dimensions including eyestrain (the speech view is easy to follow), enjoyment (the mode is fun to use), naturalness (the mode is natural to use and whether the balloons work as real-world objects), and assistance (the mode helps in the conversation) were rated as dependent variables in this task.

In the second study, participants were invited to watch the recorded speech view with the modes S, F, and M. We randomly selected a one-minute-long saved raw video clip from all clips from the first study, created its corresponding augmentation in three modes, and played it to all participants. The video clip was muted. In this experiment, our approach quantitatively recorded the eye gaze paths to check the user's attention. The dependent variables in the experiment were the average saccade velocity of the eye gaze and the average distance between the eye gaze and the speaker's face. We inspected the participants' eye gaze paths with two measurements. The first measurement was the average saccade velocity of the eye gazes v_i. The unit was the number

of pixels per second. We applied an identification by velocity threshold (IV-T) fixation filter [44] with a $15^\circ/s$ threshold. The saccade velocity measured whether the user frequently moved his eye at a considerable distance. When the participants switch between text and the speaker, they must overcome longer viewing distances. Higher saccade velocity is an essential factor in causing fatigue [45]. The second measurement was the average Euclidean distance between the eye gaze and the speaker's face $d_i = ||\mathbf{a}_i - \mathbf{q}'_{1i}||$, where $\mathbf{q}'_{1i}$ is the screen coordinate of the speaker's nose at the ith frame. The unit was the number of pixels. The distance d_i measured the deviation of the mutual eye gaze in the conversation. With word balloons close to the speaker, the audience can better focus on the image content.

2) Hypotheses: We formulate four hypotheses considering the distribution of attention and saccade changes:

H 1: Mode M is more joyful than other modes.

H 2: Mode M is more natural than other modes.

H 3: The average saccade velocity of the eye gazes for mode S is higher than others. The fixed subtitle will cause participants to perform faster saccades, resulting in more efforts and exhaustion.

H 4: The average distance between the eye gaze and the speaker's face for mode S is higher than others. The speaker's face stays at the center of the view during face-to-face conversations, resulting in larger gaze distances when reading subtitles.

H1 and H2 concern the placement and motion of word balloons. As the word balloons in MR better fit the real world than the static graphics on the screen, we hypothesize mode M is more natural and joyful. H3 and H4 are derived from previous findings [17] for videos with gaze-adaptive captions. We hypothesize that the designed word balloons in MR would result in similar behavior.

3) Procedure: We recruited 24 participants (17 male) in the first study, randomly divided into 12 pairs. The average age was 22 years old ($SD = 1.7$). For the second study, we invited 12 participants (6 male). The average age was 21 years old ($SD = 2.0$). All participants were undergraduates majoring in science, engineering, or arts. All of them had comics reading experience, and 2 had experience drawing comics or user interface design. Before each study, we gave the participants a 10-min session to explain the user interface and the task description. They were encouraged to ask any questions. The participants in the first study could freely play with our system to understand the interface of the application before the study. Furthermore, the participants in the second study were seated in front of a desktop screen (with 3480×2160 resolution). The Tobii Eye Tracker 4 C was installed on the desktop screen to acquire the eye gaze paths in this experiment. We performed the calibration of the eye tracker for each participant in the second study.

4) Results:

a) Results of the first study: The subjective feedback was shown in Fig. 8. For eye strain and assistance, as the Shapiro–Wilk normality test showed the normal distribution was violated ($p < 0.001$), we performed a Friedman test, which did not reveal any significant difference for eye strain ($p = 0.9$) and assistance ($p = 0.6$). For enjoyment and naturalness, as the Shapiro–Wilk normality test showed the normal distribution

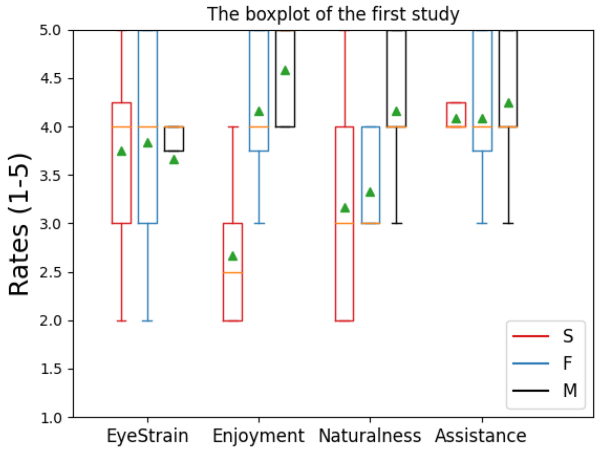


Fig. 8. Boxplot of the user study comparing the subtitle (mode S), the speech bubble [13] (mode F), and the word balloon (mode M). The green triangles shows the mean value.

TABLE III
STATISTIC OF EYE TRACKING

Mode	Mode S	Mode F	Mode M
Velocity	4574 (*SD*=499)	3797 (*SD*=491)	3710 (*SD*=424)
Distance	970.6 (*SD*=59.8)	680.2 (*SD*=49.6)	704.1 (*SD*=40.0)

was violated ($p < 0.001$), we performed a Friedman test, which revealed a significant difference for enjoyment ($p < 0.001$) and naturalness ($p = 0.018$). We also performed a Conover posthoc test for enjoyment and naturalness. For enjoyment, mode S is significantly less joyful than mode F ($p = 0.0066$) and mode M ($p = 0.0011$), and no significant difference was found between mode F and mode M ($p = 0.46$). (H1 not supported)

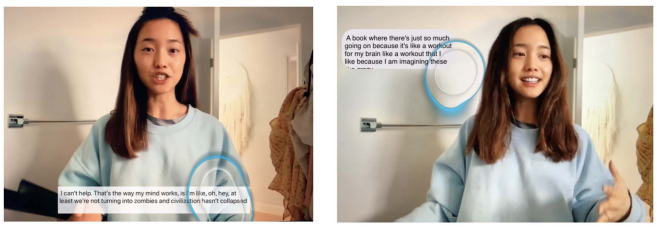
For naturalness, mode M is significantly more natural than mode S ($p = 0.017$) and mode F ($p = 0.028$), and no significant difference found between mode S and mode F ($p = 0.81$). (H2 supported)

Compared with the simple implementation in the mode F [13], the word balloons from our approach were more flexible to adapt to the conversation scene.

All participants considered word balloons to be helpful in communications under some circumstances. Most participants thought the word balloons helped them understand the conversation. We also identified two problems during the feedback section. First, four participants reported that it was not easy to look at the text on the bubble when the bubble was moving too fast. Second, the delays in speech recognition make it difficult to keep up the communication in real time.

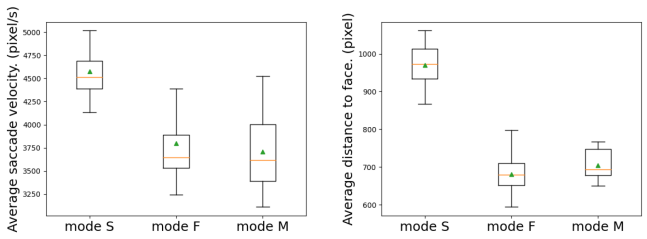
b) Results of the second study: We illustrated the eye gazes at two frames in Fig. 9. We plotted the boxplot of the two measurements in Fig. 10. In the figure, we found that with the word balloons, the eye gaze paths were smoother and better matched to the speaker's face motion compared to the subtitle mode. We also reported the average and standard deviation (*SD*) in Table III. The unit is the number of pixels per second and the number of pixels, respectively.

For the average saccade velocity of the eye gazes, as the Shapiro–Wilk normality test showed the normal distribution



(a) (b)

Fig. 9. Recorded eye gazes (depicted by the blue-boundary region) of the participant on the screen. (a) In the subtitle mode S, the participant frequently moved his gaze between the speaker's face and the subtitles. (b) In our word balloon mode M, the eye gaze stayed near the speaker's face where the word balloon was placed. Note that the balloon moved with the speaker's face.



(a) (b)

Fig. 10. (a) Boxplot of the average saccade velocity v of the eye gaze. (b) Boxplot of the distance d to the face center. The green triangles shows the mean value.

was not violated ($p > 0.05$), we performed a RM-ANOVA test, which revealed a significant difference ($F = 9.79, p < 0.001$). For the posthoc test, we performed a pairwise t-test with Bonferroni correction and found that both mode F ($p = 0.010$) and mode M ($p = 0.003$) had a significantly lower saccade velocity than mode S. (H3 supported)

For the distance to face, as the Shapiro–Wilk normality test showed the normal distribution was not violated ($p > 0.05$), we also performed an RM-ANOVA test, which revealed a significant difference ($F = 110.6, p < 0.001$). For the posthoc test, we performed a pairwise t-test with Bonferroni correction and found that both mode F ($p < 0.001$) and mode M ($p < 0.001$) had a significantly lower distance to face than mode S. (H4 supported)

No significant difference was found between mode F and mode M. However, our method (mode M) can fit the scene better with the help of the guidance field extracted from comics.

C. Applications and Discussion

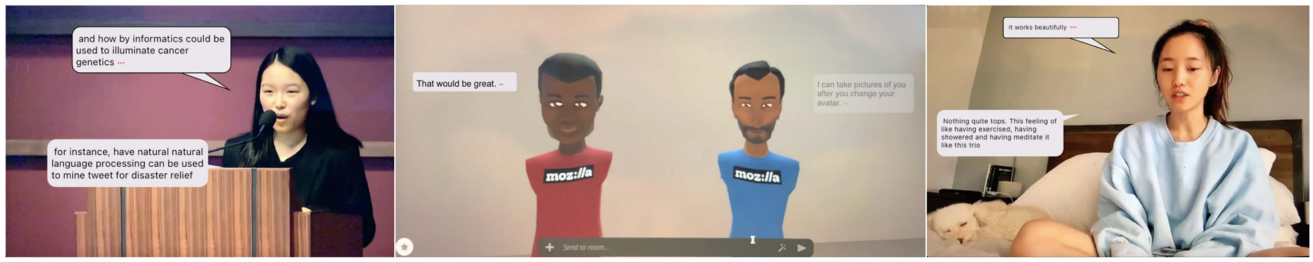
We test our approach with different videos with conversations.[1] We test our approach for scenarios including a speech, a VR meeting, and a video blog, as shown in Fig. 11. Online recognized words are displayed in the scene in real time. We see the word balloons are placed in a comic-style and smoothly update their positions in response to human motion. In Fig. 12, we show the augmented talking scenario of a captured video from Hololens2. The words are transcribed, displayed in word balloons, and attached to each speaker in the conversation. In this demo, we send the recorded video from the Hololens2 to a server, compute for the augmentation and transfer it back to the display. There is a short lag due to the computation. Because the official vision toolkits of hololens do not perfectly cover all the modules required by our system, we still use the mobile AR implementation in the user study. The social awareness issue of the mobile AR application does not affect the evaluation. We expect a mature AR application without social awareness issues with advanced tools for the developers of eyeglass AR. Our method also works for social VR scenes, as shown in the middle of Fig. 11. We convert the speech into social VR and attach the synthesized word balloons to the virtual avatars. It makes the conversation in VR feasible for broader audiences.

Based on the results above, we verify our design rationale and choice. Conventionally, subtitles are used to present sentences in conversation. Subtitles are shown either at the bottom of the view (Design B) or at a constant distance from the speaker (Design S). Both methods work in the image space. Design B is the most common way to display subtitles. As pointed out by [17], Design B is welcomed by experienced users, while Design S is preferred if the users have less experience reading subtitles. The critical feature of Design B is that the words are always displayed at a fixed position regardless of the scene. The advantage of Design S is at the cost of scene understanding, which may not always work well. However, our approach only requires a rough head pose estimation compared with other VR/AR applications. Therefore, we propose to use this dynamic layout rather than static subtitles at the bottom.

In our design, the word balloons also move in response to the 3-D motion of the speaker. It could be regarded as a VR-related feature, as the alignment of the 3-D information to the captured 2-D video will enable audiences to see the synthesized word balloons immersively. Another difference in our approach is in the image space update. Our approach learned the placement style from comics, while design S places the word balloons at a constant location concerning the detected face in the image. Our motivation is to transfer the virtual world layout style from comics to the real world to augment conversation. Because there are no strict and precise rules about word balloon placement, we choose to learn the rules from comics directly. Nevertheless, other real-world datasets should be able to mix into the existing dataset as long as they are annotated carefully. The user study results verify the advantages of these two differences in comparison with Design S.

We also prototype more extensions. Because the word balloons are virtual elements added to the screen, we can manipulate them in the augmented conversation. As shown in Fig. 13(a), the user observes an important message. By tapping, dragging, and pinning the word balloon corresponding to the message, the word balloon is kept without fading away. They are also allowed to unpin the word balloons if needed. In Fig. 13(b), we apply our approach to a doctor visit. By matching with a pharmacy library as keywords, we detect drug names and visually highlight them for more efficient communication. The appearance of the word balloons can also be extended in our framework. For example,

[1]In this work, we use the videos courtesy of online users Unspeakable, Music by Blanks, Jenn Im, Flying The Nest, Stanford, UCLA Health, Webex, Kharma Medic, Harmonizacja Kwantowa, Bozbe.



(a) (b) (c)

Fig. 11. Our approach augments different situations with word balloons (a) Speech. (b) VR meeting. (c) Vlog.

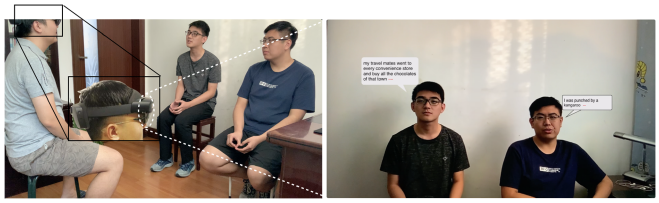
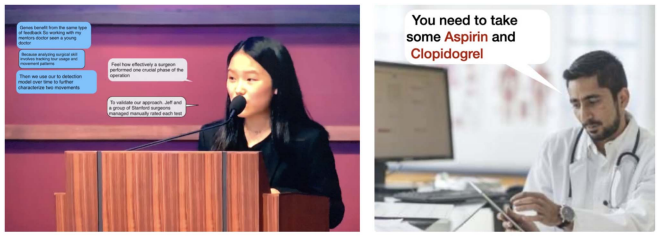
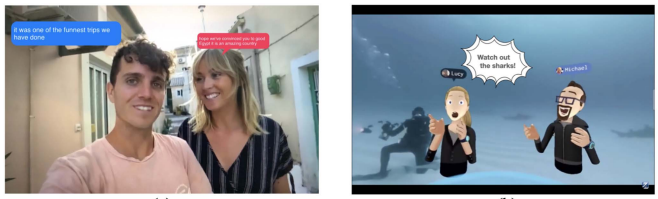


Fig. 12. We augmented the video captured from the Hololens2 in a conversation.



(a) (b)

Fig. 13. (a) Users may click and pin the word balloons containing important messages that they want to preserve. Those balloons are shown with a blue background color. (b) Keywords in the conversation may be highlighted with a different color.



(a) (b)

Fig. 14. Extensions of our approach. (a) Speaker's identity is recognized and mapped to the color of the word balloons. (b) Voice volume is mapped to the shape of a word balloon.

we can detect the speaker's identity and map it to the color of the background, as shown in Fig. 14(a). In Fig. 14(b), we set a spiky word balloon and allow a smooth transition to form a shape space of the balloons. Then we detect the volume of the voice and map it to the shape of the balloon. In this case, we may see the volume or the implied emotion by observing the shape of the word balloon.

VI. Conclusion

We introduce comic-style conversation in MR. By exploiting MR features and learning from comics, we design the variables in a template space and update them in both the object and image space. The word balloons are intelligent in that they are aware of the speaker's motion in the 3-D real world, the background information, and the speech in the conversation. Experiments show that our design is helpful in the assistance of interpersonal conversation.

A. Limitations

We assume the detected speech is associated with one speaker in the scene. If no face is detected, we initialize the word balloon in the center of the view. However, there are cases where the speaker is outside the view of face-to-face communication in an open world. Correctly showing the recognized words in balloons and attaching them requires locating the speaker with only the audio information. In this case, additional hardware, such as microphone arrays, may be required. Our MR approach's quality depends on visual and speech data processing. Although speech-to-text, speech segmentation, and speaker recognition for a prerecorded audio track have been studied for years, obtaining the texts with punctuation from streaming audio stably is still challenging. The speech processing module may fail if the speaker speaks very fast or has a strong accent. Many people speaking at the same time may also interrupt our approach. We expect our approach to be more helpful with the development of speech-processing techniques in the future.

B. Future Work

Face-to-face communication contains rich multimodal signals. In the future, we will study how to extract more useful information in a face-to-face conversation and explore how to map the high-dimensional information to the design space of a word balloon, including the space of font, text colors, highlights, and more shape design of the balloon. So far, we only automate the placement and resizing of the word balloons. Users customize the other dimensions of the word balloon design. We will also explore using other commercially available natural language processing tools to facilitate face-to-face communication. Adapting our approach to a collaborative MR environment would be another direction for our future work. In this work, we

show that the features learned from the comics can be used to support real-world communication. It gives us another hint that the map between the natural world and the imaginary world can be estimated with computational tools. The map bridges the real and imaginary world and is helpful for MR. It will be interesting to study this map further to assist more tasks in the real-world with MR techniques.

REFERENCES

[1] D. Kurlander, T. Skelly, and D. Salesin, "Comic chat," in *Proc. 23rd Annu. Conf. Comput. Graph. Interactive Techn.*, 1996, pp. 225–236.

[2] S. McCloud, *Understanding Comics: The Invisible Art*. Northampton, MA, USA: Tundra Pub., 1993.

[3] P. Milgram and F. Kishino, "A taxonomy of mixed reality visual displays," *IEICE Trans. Inf. Syst.*, vol. 12, pp. 1321–1329, 1994.

[4] M. Speicher, B. D. Hall, and M. Nebeling, "What is mixed reality?," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2019, pp. 1–15.

[5] O. Augereau, M. Iwata, and K. Kise, "An overview of comics research in computer science," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit.*, 2017, vol. 3, pp. 54–59.

[6] T.-T. Wong, T. Igarashi, Y.-Q. Xu, and D. Shi, "Computational manga and anime," in *Proc. SIGGRAPH Asia Courses*, 2013, pp. 1–52.

[7] B.-K. Chun, D.-S. Ryu, W.-I. Hwang, and H.-G. Cho, "An automated procedure for word balloon placement in cinema comics," in *Proc. Adv. Vis. Comput.*, 2006, pp. 576–585.

[8] M. Toyoura, T. Sawada, M. Kunihiro, and X. Mao, "Using eye-tracking data for automatic film comic creation," in *Proc. Symp. Eye Tracking Res. Appl.*, 2012, pp. 373–376.

[9] Y. Cao, R. W. H. Lau, and A. B. Chan, "Look over here: Attention-directing composition of manga elements," *ACM Trans. Graph.*, vol. 33, no. 4, pp. 1–11, 2014.

[10] T. Chen, P. Tan, L.-Q. Ma, M.-M. Cheng, A. Shamir, and S.-M. Hu, "Poseshop: Human image database construction and personalized content synthesis," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 5, pp. 824–837, May 2013.

[11] G. Jing, Y. Hu, Y. Guo, Y. Yu, and W. Wang, "Content-aware video2comics with manga-style layout," *IEEE Trans. Multimedia*, vol. 17, no. 12, pp. 2122–2133, Dec. 2015.

[12] W.-I. Hwang, P.-J. Lee, B.-K. Chun, D.-S. Ryu, and H.-G. Cho, "Cinema comics: Cartoon generation from video stream," in *Proc. GRAPP*, 2006, pp. 299–304.

[13] Y.-H. Peng et al., "Speechbubbles: Enhancing captioning experiences for deaf and hard-of-hearing people in group conversations," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2018, pp. 1–10.

[14] T. Kurahashi, R. Sakuma, K. Zempo, K. Mizutani, and N. Wakatsuki, "Retrospective speech balloons on speech-visible ar via head-mounted display," in *Proc. IEEE Int. Symp. Mixed Reality Augmented Reality*, 2018, pp. 423–424.

[15] R. S. Kushalnagar, G. W. Behm, A. W. Kelstone, and S. Ali, "Tracked speech-to-text display: Enhancing accessibility and readability of real-time speech-to-text," in *Proc. 17th Int. ACM SIGACCESS Conf. Comput. Accessibility*, 2015, pp. 223–230.

[16] S. Rothe, K. Tran, and H. Hußmann, "Dynamic subtitles in cinematic virtual reality," in *Proc. ACM Int. Conf. Interactive Experiences TV Online Video*, 2018, pp. 209–214.

[17] K. Kurzhals, F. Göbel, K. Angerbauer, M. Sedlmair, and M. Raubal, "A view on the viewer: Gaze-adaptive captions for videos," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2020, pp. 1–12.

[18] L. Sidenmark, N. Kiefer, and H. Gellersen, "Subtitles in interactive virtual reality: Using gaze to address depth conflicts," in *Proc. Workshop Emerg. Novel Input Devices Interaction Techn.*, 2019, pp. 1–6.

[19] B. Bell, S. Feiner, and T. Hollerer, "View management for virtual and augmented reality," in *Proc. ACM Symp. User Interface Softw. Technol.*, 2001, pp. 101–110.

[20] J. Orlosky, K. Kiyokawa, T. Toyama, and D. Sonntag, "Halo content: Context-aware viewspace management for non-invasive augmented reality," in *Proc. 20th Int. Conf. Intell. User Interfaces*, 2015, pp. 369–373.

[21] S. Pick, B. Hentschel, I. Tedjo-Palczynski, M. Wolter, and T. Kuhlen, "Automated positioning of annotations in immersive virtual environments," in *Proc. Joint Virtual Reality Conf.*, 2010, pp. 1–8.

[22] R. Grasset, T. Langlotz, D. Kalkofen, M. Tatzgern, and D. Schmalstieg, "Image-driven view management for augmented reality browsers," in *Proc. IEEE Int. Symp. Mixed Augmented Reality*, 2012, pp. 177–186.

[23] M. Tatzgern, D. Kalkofen, R. Grasset, and D. Schmalstieg, "Hedgehog labeling: View management techniques for external labels in 3-D space," in *Proc. IEEE Virtual Reality*, 2014, pp. 27–32.

[24] J. B. Madsen, M. Tatzqern, C. B. Madsen, D. Schmalstieg, and D. Kalkofen, "Temporal coherence strategies for augmented reality labeling," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 4, pp. 1415–1423, Apr. 2016.

[25] Z. Chen, J. Li, Y. Hua, R. Shen, and A. Basu, "Multimodal interaction in augmented reality," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 2017, pp. 206–209.

[26] F. Wang, H. Nagano, K. Kashino, and T. Igarashi, "Visualizing video sounds with sound word animation to enrich user experience," *IEEE Trans. Multimedia*, vol. 19, no. 2, pp. 418–429, Feb. 2017.

[27] K. Perlin, "Future reality: How emerging technologies will change language itself," *IEEE Comput. Graph. Appl.*, vol. 36, no. 3, pp. 84–89, May/Jun. 2016.

[28] R. Mayer, "Using multimedia for e-learning," *J. Comput. Assist. Learn.*, vol. 33, no. 5, pp. 403–423, 2017.

[29] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2 d 3d face alignment problem? (and a dataset of 230,000 3-D facial landmarks)," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1021–1030.

[30] X. Gao, X. Hou, J. Tang, and H. Cheng, "Complete solution classification for the perspective-three-point problem," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 8, pp. 930–943, Aug. 2003.

[31] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1867–1874.

[32] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[33] A. Fujimoto, T. Ogawa, K. Yamamoto, Y. Matsui, T. Yamasaki, and K. Aizawa, "Manga109 dataset and creation of metadata," in *Proc. 1st Int. Workshop Comics Anal., Process. Understanding*, 2016, pp. 1–5.

[34] M. Xie, M. Xia, X. Liu, C. Li, and T.-T. Wong, "Seamless manga inpainting with semantics awareness," *ACM Trans. Graph.*, vol. 40, no. 4, pp. 96:1–96:11, Aug. 2021.

[35] D. Anonymous community and G. Branwen, "Danbooru2019: A large-scale crowdsourced and tagged anime illustration dataset," Jan. 2020, Accessed: Dec. 2019. [Online]. Available: https://www.gwern.net/Danbooru2019

[36] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[37] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[38] C. W. Reynolds, "Flocks, herds and schools: A distributed behavioral model," in *Proc. 14th Annu. Conf. Comput. Graph. Interactive Techn.*, 1987, pp. 25–34.

[39] iFLYTEK, "iflytek online text to speech," 2020. [Online]. Available: https://www.xfyun.cn/services/rtasr

[40] Google, "Google cloud text-to-speech," 2020. [Online]. Available: https://cloud.google.com/text-to-speech

[41] S. O. Arik et al., "Deep voice: Real-time neural text-to-speech," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 195–204.

[42] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, "Utterance-level aggregation for speaker recognition in the wild," in *Proc. Int. Conf. Acoust., Speech, Signal Process.* (Oral), 2019, pp. 5791–5795.

[43] Y. Hu, J. Kautz, Y. Yu, and W. Wang, "Speaker-following video subtitles," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 11, no. 2, pp. 1–17, 2015.

[44] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Proc. Symp. Eye Tracking Res. Appl.*, 2000, pp. 71–78.

[45] A. Bahill and L. Stark, "Overlapping saccades and glissades are produced by fatigue in the saccadic eye movement system," *Exp. Neurol.*, vol. 48, no. 1, pp. 95–106, 1975.