Modeling and Improving Text Stability in Live Captions

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

Jun Zhang
Google Research
Mountain View, CA, USA
xiaomu@google.com

Leonardo Ferrer
Google Research
Mountain View, CA, USA
leoferrer@google.com

Susan Xu
Google Research
San Francisco, CA, USA
xsusan@google.com

Vikas Bahirwani
Google Research
San Francisco, CA, USA
vikasbahirwani@google.com

Boris Smus
Google Research
Seattle, WA, USA
smus@google.com

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

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

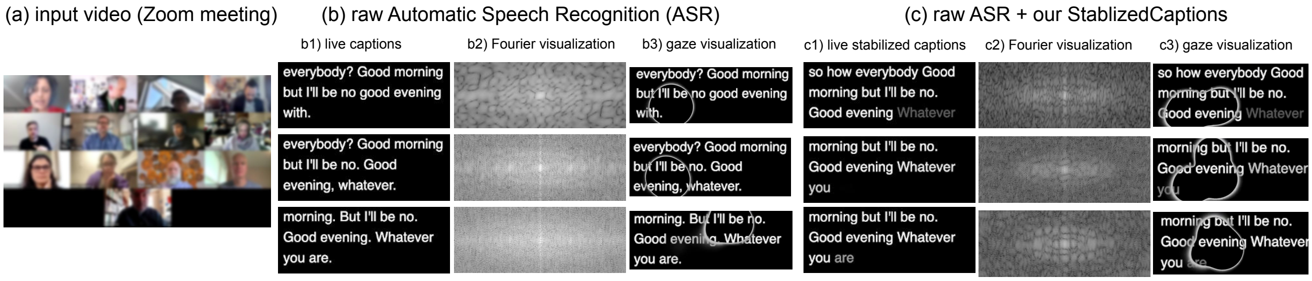

(a) input video (Zoom meeting)
(b) raw Automatic Speech Recognition (ASR)
(c) raw ASR + our StablizedCaptions
b1) live captions
b2) Fourier visualization
b3) gaze visualization
c1) live stabilized captions
c2) Fourier visualization
c3) gaze visualization
everybody? Good morning but I'll be no good evening with.
everybody? Good morning but I'll be no. Good evening, whatever.
morning. But I'll be no. Good evening. Whatever you are.
everybody? Good morning but I'll be no good evening with.
everybody? Good morning but I'll be no. Good evening, whatever.
morning. But I'll be no. Good evening. Whatever you are.
so how everybody Good morning but I'll be no. Good evening Whatever
morning but I'll be no. Good evening Whatever you
morning but I'll be no. Good evening Whatever you are
so how everybody Good morning but I'll be no. Good evening Whatever
morning but I'll be no. Good evening Whatever you
morning but I'll be no. Good evening Whatever you are

Figure 1: Visualization of the stability issues in live captions. (a) shows an example input video of a public Zoom meeting, (b) shows results of raw ASR layout with b1) instable live captions, b2) visualization of Discrete Fourier Transform (DFT) between frames, b3) visualization of user's gaze (c) shows results of our stabilized live captions with c1) improved layout, c2) visualization of DFT, and c3) visualization of gaze.

ABSTRACT

In recent years, live captions have gained significant popularity through its availability in remote video conferences, mobile applications, and the web. Unlike preprocessed subtitles, live captions require real-time responsiveness by showing interim speech-to-text results. As the prediction confidence changes, the captions may update, leading to visual instability that interferes with the user's viewing experience. In this paper, we characterize the stability of live captions by proposing a vision-based flickering metric using luminance contrast and Discrete Fourier Transform. Additionally, we assess the effect of unstable captions on the viewer through task load index surveys. Our analysis reveals significant correlations between the viewer's experience and our proposed quantitative metric. To enhance the stability of live captions without compromising responsiveness, we propose the use of tokenized alignment, word updates with semantic similarity, and smooth animation. Results from a crowdsourced study (N=123), comparing four strategies, indicate that our stabilization algorithms lead to a significant reduction in viewer distraction and fatigue, while increasing viewers' reading comfort.

CCS CONCEPTS

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

KEYWORDS

Live Captions; Real-Time Transcription; Visual Instability; Flickering Metric; Speech-to-text; Text Stability;

ACM Reference Format:
Xingyu “Bruce” Liu, Jun Zhang, Leonardo Ferrer, Susan Xu, Vikas Bahirwani, Boris Smus, Alex Olwal, and Ruofei Du. 2023. Modeling and Improving Text Stability in Live Captions. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA

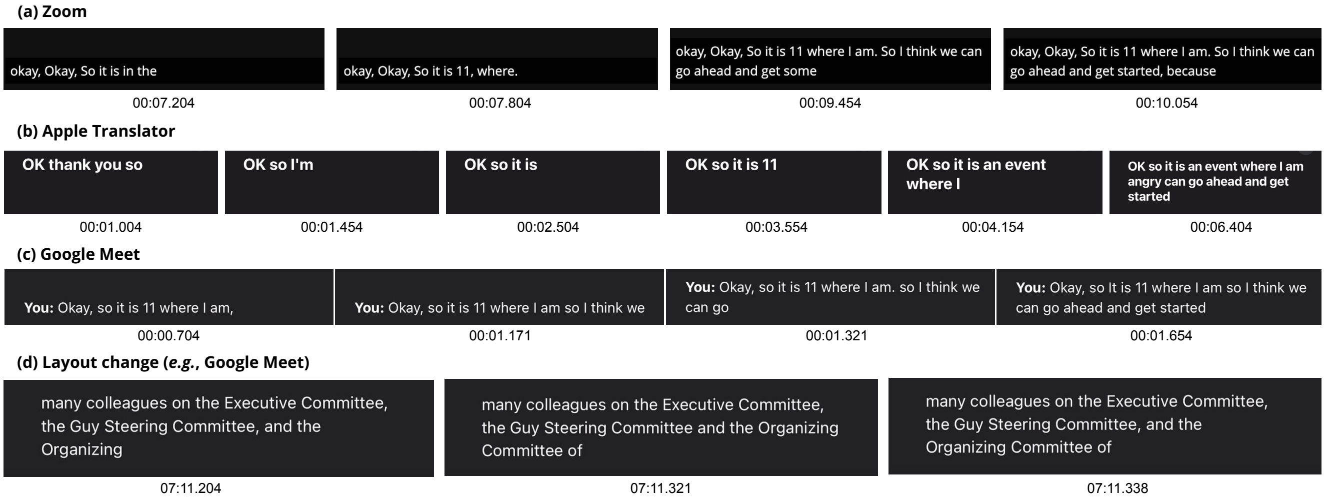
1 INTRODUCTION

The widespread adoption of Automatic speech recognition (ASR) technology has made conversations more accessible with live captions in remote conferencing software (*e.g.*, Google Meet, Zoom, Microsoft Teams), mobile applications (*e.g.*, Live Transcribe & Notifications in Android, Translate in iOS), as well as head-worn displays [8, 27]. However, to maintain real-time responsiveness, live caption systems often display interim ASR predictions that are updated as new utterances are received and confidence changes, as shown in Figure 1. (Figure 4b illustrates how the Apple Translator

*Corresponding author

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

ACM ISBN 978-1-4503-9422-2/23/04.
https://doi.org/10.1145/3544549.3585609


(a) Zoom

okay, Okay, So it is in the
00:07.204

okay, Okay, So it is 11, where.
00:07.804

okay, Okay, So it is 11 where I am. So I think we can go ahead and get some
00:09.454

okay, Okay, So it is 11 where I am. So I think we can go ahead and get started, because
00:10.054

(b) Apple Translator

OK thank you so
00:01.004

OK so I'm
00:01.454

OK so it is
00:02.504

OK so it is 11
00:03.554

OK so it is an event where I
00:04.154

OK so it is an event where I am angry can go ahead and get started
00:06.404

(c) Google Meet

You: Okay, so it is 11 where I am,
00:00.704

You: Okay, so it is 11 where I am so I think we
00:01.171

You: Okay, so it is 11 where I am. so I think we can go
00:01.321

You: Okay, so It is 11 where I am so I think we can go ahead and get started
00:01.654

(d) Layout change (*e.g.*, Google Meet)

many colleagues on the Executive Committee, the Guy Steering Committee, and the Organizing
07:11.204

many colleagues on the Executive Committee, the Guy Steering Committee and the Organizing Committee of
07:11.321

many colleagues on the Executive Committee, the Guy Steering Committee, and the Organizing Committee of
07:11.338

Figure 2: Key frames and associated timestamps which demonstrates the common stability issues in the current generation of commercial live caption systems. (a) shows four sampled captions in Zoom: "in the" is replaced by "11, where."; "some" is replaced by "get started, because". (b) shows six sampled captions in Apple Translator: "thank you" is removed and "so" is shifted to the prior position. "I'm" is replaced by "it is", font size becomes smaller when more transcription is presented. (c) shows four sampled captions in Google Meet: the comma after "I am" is removed and then added as a period; "it" is replaced by "It". (d) show a common layout change issue across all commercial software: "Organizing" is moved to the prior line with a shorter sentence beforehand and moved to the next line with a longer sentence. Screenshots are captured with the same YouTube video [34] played in a quiet room with the latest software in August 2022. Please refer to the supplementary video for the full recordings.

iteratively corrects intermediate ASR results (*"OK thank you so", "Ok so I'm", "OK so it is 11"*) until it obtains enough context to finalize the correct transcription (*"OK so it is is an event"*). Such text instability in live captions may significantly impair users' reading experience. Users are often distracted by changes in layout, modification of words, and adjustment of punctuation in live captions, as evidenced by Heunerfauth's work [3] and our formative studies.

In this work, we formalize the problem of text stability in live captioning through the following contributions: **(1) a quantitative metric to model the stability of live captions**. Specifically, we explored a vision-based flicker metric using luminance contrast and Discrete Fourier Transform; **(2) an algorithm to stabilize the rendering of live captions**, via tokenized alignment, semantic merging, and smooth animation. We compared our approach with raw ASR results and different strategies; and **(3) a crowdsourced study (N=123) to understand viewers' experience with live captioning**. We asked crowdworkers to watch videos and their accompanying captions with different stability strategies, and rate their agreement with six statements including the level of comfort, distraction, fatigue, etc. Our statistical analysis demonstrates a strong correlation between our proposed flickering metric and viewers' experience, and that our proposed stabilization techniques significantly improved viewers' experience.

2 RELATED WORK

2.1 Challenges in using ASR systems

While Automatic Speech Recognition (ASR) systems have been widely adopted for services like voice assistants, computer-mediated communication tools and assistive applications [2, 20, 21], ASR is, by nature, imperfect. This has led to concerns over transcription quality. Butler *et al.* and Mirzaei *et al.* [4, 24] found that while d/Deaf or Hard of Hearing (DHH) students found the benefits of ASR, they identify accuracy and readability of ASR as major pain points, and struggle with low-quality captions [3, 11]. Some researchers argue that understanding ASR-generated captions requires higher cognitive effort than errors produced by human captioners [16]. There is a hesitation to adopt ASR if it means that high-quality professional captioning is replaced with lower-quality automated solutions [25, 29].

2.2 Existing approaches to address ASR errors

Researchers have attempted to address ASR errors by adjusting the appearance of captions to reflect uncertainty. For example, Berke *et al.* [3] conducted studies with 128 DHH participants to display ASR word confidence along with the transcription, but their participants found it distracting. Other researchers also explored ways of displaying recognition results that take into account the confidence of each word in the transcription, such as color-coding and bolding [28, 33].

Additionally, studies have been conducted on enhancing comprehension through transcription formatting. Kafle *et al.* [10] found that videos with captions containing highlighted words had lower perceived task-load ratings and were easier to read and follow by DHH participants. They also found that DHH participants preferred boldface, word-level highlighting for captions in the educational lecture video genre [9]. Other studies, like Hong *et al.* [7], considered how to highlight video transcriptions to align them with the speech signal and illustrate voice volume. Angerbauer *et al.* [1] explored compressing subtitles for enhanced readability.

However, the varying degrees of ASR confidence can still result in visual jitters in live streaming scenarios, as seen in commercial ASR services [18, 32]. Researchers have attempted to address this issue by analyzing different types of ASR flicker and the impact of model training techniques on stability [32]. Li *et al.* [19] applied an encoder states revision strategy to improve the quality of causal models, which have low latency but worse quality than traditional look-ahead ASR models. There is also substantial work in the NLP and MT space to improve text stability from the translation side, often at some cost to latency [22, 26, 39].

Given these intermediate results of commercial ASR systems and their error patterns, we believe that there is an opportunity to systematically model visual jitter in live captions, and investigate strategies to stabilize the text rendering.

2.3 Evaluating Offline and Live Captions

Word Error Rate (WER) is one of the most commonly used metrics in evaluating the quality of Speech Recognition. However, WER focuses on only on spoken word sequences, excluding factors such as delay, positioning, punctuation, which are included in subsequent metrics such as NER [31]. In terms for real-time captions, prior research has performed empirical studies with DHH users [29]. Researchers have used open-ended questions [23], scalar-response questions [15, 17, 35], and comprehension questions [3, 6, 36]. In terms of user behaviors, researchers also used eye tracking methods to quantitatively evaluate users' attention of subtitles of different layouts [14] and to measure cognitive load [12, 13]. By recording users' eye tracking data, Kurzhals *et al.* were able to apply established metrics such as fixation count and saccade length, which are known to be important factors in eye strain.

Our work focuses on the problem of live captions being updated at a high frequency with prior words modified in real-world scenarios, rather than in an offline system where prior words are not modified. We take inspiration from Winkler *et al.* in the computer vision domain who studied the problem of temporal flicker in watermarking algorithms, which is analogous to the issue of caption flicker caused by ASR instability in our work.

3 A QUANTITATIVE METRIC FOR MODELING TEXT STABILITY

The dynamic nature of Automatic Speech Recognition (ASR) predictions has led to increasing text instability, which can negatively impact user experience and comprehension. To objectively evaluate the performance of live captioning systems, a quantitative metric for text stability is essential for guiding the development of algorithms that enhance live captions. In this regard, we propose adapting the perceptual metrics proposed by Winkler *et al.* to measure text stability. This metric offers a quantitative measure of the visibility of changes between consecutive frames, providing a valuable insight into the visibility of changes in caption text over time.

We illustrate our adaption of the flicker metrics to measure text stability in Figure 3. Given a grayscale live caption video at the resolution of $W \times H$, we compute the the difference in luminance between adjacent frames $\mathbf{I}_i$ and $\mathbf{I}_{i-1}$ to deltas $\mathbf{D}_i = |\mathbf{I}_i - \mathbf{I}_{i-1}|, i \in \{2, \cdots, N\}$. We discard subtle luminance changes by setting $\mathbf{D}_i$ to 0 when it is below 50, such that fading animations do not yield a large metric. By averaging the flickering across all consecutive frames, we define the metric of luminance contrast as $\mathcal{M}_{\text{luminance}} = \frac{1}{N-1} \sum_{i=2}^{N} \overline{\mathbf{D}_i}$.

Next, we apply the discrete Fourier transform to the deltas, which gives a vector of Fourier coefficients $\mathbf{c}_i = \text{DFT}(\mathbf{D}_i)$. We then compute a sum s_L over low frequencies and a sum s_H over high frequencies to get a per-frame flicker $s_L + s_H$. Finally, we average the flickering across all consecutive frames in the video to get a per-video flicker value:

$$s_L(i) = \frac{1}{f_M - f_L} \sum_{k=f_L}^{f_M} \mathbf{c}_k(i),$$

$$s_H(i) = \frac{1}{f_H - f_M} \sum_{k=f_M}^{f_H} \mathbf{c}_k(i),$$

$$\mathcal{M}_{\text{flicker}} = \frac{1}{(N-1) \cdot m} \sum_{i=2}^{N} \left(s_L(i) + s_H(i)\right),$$

where N is the total number of frames, $\{\mathbf{I}_i\}_{i=1}^{N}$ are frames of the input video, DFT represents computing Fourier coefficients, and f_L, f_M, f_H are predefined frequency limits. As in Winkler *et al.* [38], we use frequency limits of $f_L = 1\%, f_M = 16\%$, and $f_H = 80\%$ relative to the maximum frequency and apply a normalization factor m, which is detailed in the Appendix. Please refer to the supplementary material for the Python code.

4 STABILIZED CAPTIONS

To improve the stability of live captions, we propose an algorithm that performs tokenized alignment, semantic merging, and smooth animation, which takes as input a sequence of interim ASR predictions and outputs a stabilized text string.

4.1 Problem Definition

Given D (old) words $\{w_i\}, i < D, i \in \mathbb{N}$ already rendered in the live captions with line breaks, the system receives E (new) words $\{v_j\}, j < E, j \in \mathbb{N}$, and expects a stabilized text string with line breaks given a fixed screen resolution.

For each incoming sentence, we first convert it into an array of tokens. Each token contains a tuple of the raw representation (original word) and the filtered representation (lower-case, without punctuations). Given an old sequence $\boldsymbol{X}$ consisting of tokens $\{\boldsymbol{X}_i, i < |\boldsymbol{X}|, i \in \mathbb{N}\}$ and a newly updated sequence $\boldsymbol{Y}$ consisting of tokens $\left\{\boldsymbol{Y}_j, j < |\boldsymbol{Y}|, j \in \mathbb{N}\right\}$, our goal is to determine the set of words to render with the new words update. A baseline approach would be to fully trust the speech-to-text model and directly render sequence $\boldsymbol{Y}$. However, this would often introduce instability in the live captions as shown in Figure 2. A possible solution is to leverage the confidence or stability score in the real-time speech-to-text engines. However, according to prior art [37] and our own experimentation, the reliability of the confidence score may vary, and often leads to added delay in the rendering of live captions. For instance, in Google Cloud's speech-to-text API, the confidence scores are calculated by aggregating the "likelihood" values assigned to

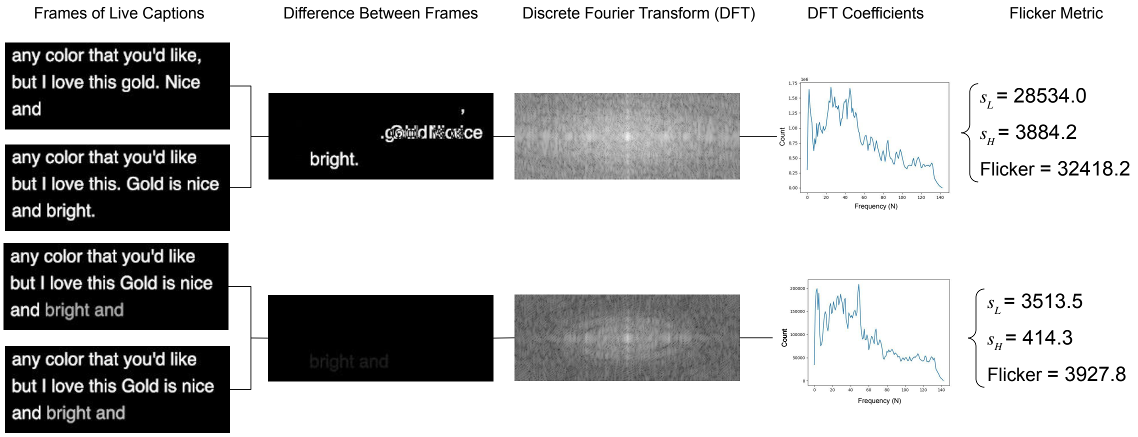

Frames of Live Captions
Difference Between Frames
Discrete Fourier Transform (DFT)
DFT Coefficients
Flicker Metric
any color that you'd like, but I love this gold. Nice and
any color that you'd like but I love this. Gold is nice and bright.
bright.
Frequency (N)
Count
s_L = 28534.0
s_H = 3884.2
Flicker = 32418.2
any color that you'd like but I love this Gold is nice and bright and
any color that you'd like but I love this Gold is nice and bright and
bright and
s_L = 3513.5
s_H = 414.3
Flicker = 3927.8

Figure 3: Illustration of the flicker metric for measuring text stability using Discrete Fourier Transform.

each word in the input audio, and provided only for results where `is_final=true`, *i.e.*, not interim.

Nevertheless, our algorithm is versatile and works with or without a confidence score — we focus on the *rendering* strategies of live captions, such as alignment, merging, and animation, to improve text stability. When a reliable confidence score is present, one can further refine the intermediate results by applying a threshold or incorporating it into the semantic-similarity score.

As shown in Figure 4, in the process of rendering live captions from sequences of ASR predictions, we identify three cases of changes when merging $\boldsymbol{X}$ (Old) and $\boldsymbol{Y}$ (New): For **Case A** (*addition of tokens to the end of captions*), there are newly updated, incoming tokens "How about" at the end of the original speech that is not going to match any old tokens. For **Case B** (*addition / deletion of tokens, not at the end of captions*), in B1, there are new words "I" and "friends" added within the laid out tokens. Here, "I" may or may out impact the overall comprehension of the caption, but it may lead to layout change as shown in Figure 4(d). Such layout changes are not desired in a live captions scenario as it often moves positions of a large chunks of texts on screen, causing significant jitter and users have to spend extra time to scan the screen and find the current focus. Also, the addition of "friends" affects the meaning of the sentence, so it is important to ensure that this change can be properly incorporated into the captions. In B2, "in" is removed from the newly updated sentence. For **Case C** (*re-caption of tokens*), in C1, "disney land" is to be updated to "Disneyland", however, this does not impact the overall semantics of the sentence. In C2, "it's" is updated to "It was" and it is typically upon user's preferences whether or not to update this.

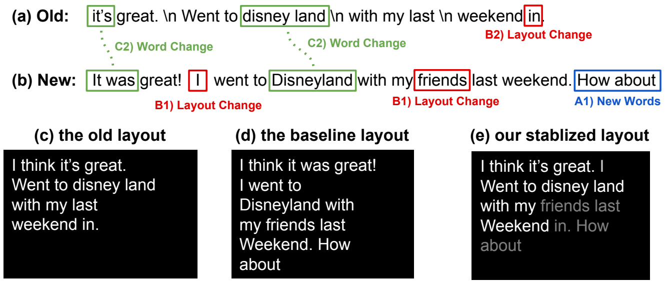


Figure 4: An example of the instable caption merging issue. Given an old sequence of words (a) with line breaks laid out in (c) and a new sequence of words (b), the baseline approach directly update the layout with the new text, regardless of their importance while we propose (e), a stabilized approach that considers existing line breaks and words and only update words that are semantically different (*e.g.*, the new "I" and "friends". Similar phrases like "disney land" and "Disneyland" are not changed.

Thus, to maximize text stability, our goal is to align the old sequence $\boldsymbol{X}$ with the new sequence $\boldsymbol{Y}$ and update them in a way that makes minimal changes to the existing layout while ensuring accurate and meaningful captions.

4.2 Caption Alignment and Merging

We leveraged a variant of the Needleman-Wunsch algorithm with dynamic programming to align two sequences. We allow a maximum skipping of $K = 3$ words when matching the sequences and always allow the algorithm to skip line breaks, *i.e.*, the line breaks are preserved in their original place after merging. With this alignment algorithm, we obtain actions as illustrated as cases A, B, and C in Figure 4. During the merging stage, we directly add Case A tokens, and add line breaks as needed by measuring if the current line exceeded 95% of the maximum width of the line. This leaves extra padding at the end of each line to allow additional punctuation and plural forms ("s") added to each line. For Case B tokens, we only add them if their addition does not break an existing line. A more effective strategy would be to incorporate them if their semantic importance surpasses a particular threshold. For Case C tokens, we compare their semantic similarity and only update them if they are semantically different and will not cause new line breaks.

4.3 Semantic-similarity Aware Updates

Our algorithm depends on a semantic similarity oracle to update the Case C tokens. We first removed all the English stop-words and punctuations, and used NLTK's WordNetLemmatizer[1] to lemmatize the remaining words and transformed all the words to lowercase.

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

We then map the original and updated words into a vector space using SentenceTransformers [30], and measure their semantic similarity by computing the dot product of the two vectors. We consider two tokens semantically similar if their similarity score is greater than 0.85.

4.4 Smooth Animation (scrolling and fading)

Finally, we leverage animations to guide people's fixations and reduce visual jitter. Hence we implement smooth scrolling and fading in of newly added tokens to further stabilize the overall layout of the live captions.

5 EVALUATION: CROWD-SOURCED STUDY

We conducted a crowdsourced study with 123 participants to (1) examine the correlation of our proposed flicker metric with users' experience with live captions, and (2) demonstrate the effectiveness of our stabilization techniques.

5.1 Dataset

We manually selected 20 videos in YouTube to obtain a broad coverage of use cases. This sample contains videos from different categories including video conferences, documentaries, academic talks, tutorials, news, comedy, and others. For each video, we selected a 30-second clip, with at least 90% speech.

5.2 Conditions

We prepared four types of renderings of live captions to compare:

C0 **Raw ASR**: raw speech-to-text results from Google's cloud speech-to-text API.
C1 **Raw ASR + thresholding**: only display interim speech-to-text result if its confidence score is higher than 0.85.
C2 **Stabilized Captions**: our stabilized approach as introduced in Section 4.
C3 **Stabilized Captions Smooth**: stabilized captions with smooth animation (scrolling + fading).

We computed flicker metrics of captions for all conditions of the 20 video samples using the algorithm in section 3.

5.3 Procedure

We collected per-condition text stability ratings by sending the collected sentences to 123 crowdworkers recruited from Amazon Mechanical Turk (MTurk). We first generated a video recording of live captions of each video in our dataset for all conditions. Crowdworkers were asked to watch the recorded live captions, and rate their agreement with six statements on a Likert scale from 1 – *Strongly Disagree* to 7 – *Strongly Agree*. Statements were adopted from questionnaires used in previous eye-tracking research on video captions [5, 14].

Q1 **Comfort:** The live captions are comfortable to read.
Q2 **Distraction:** The live captions were distracting.
Q3 **Easy to read:** The captions were easy to read and follow.
Q4 **Easy to follow video:** I can easily follow the video content.
Q5 **Fatigue:** I felt eye fatigue or eye tiredness reading the live captions.
Q6 **Impaired Experience:** The live captions impaired my viewing experience with the video.

Behavioral Measurement	Correlation to Flickering Metric
Qualitative Ratings	
Comfort	-0.294^{***}
Distraction	0.328^{***}
Easy to read	-0.305^{***}
Easy to follow video	-0.286^{***}
Fatigue	0.361^{***}
Impaired Experience	0.307^{***}
Note:	$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Table 1: Spearman correlation tests of our proposed flickering metric to users' reported behaviroal measurements in our user studies.

Each task was rated by three different crowdworkers, and workers were allowed to work on multiple HITs (Human Intelligence Tasks). We selected workers to be within the United States and have a history approval rate beyond 95%. We paid crowdworkers $0.35 for each completed HIT. We also asked crowdworkers to enter the last word in the captions as an attention check, and removed and republished the answers that did not match the last word in the task.

5.4 Results

5.4.1 Correlation between flicker metric and user experience. Through Spearman's Correlation tests (Table 1), our flicker metric is demonstrated to have statistically significant ($p < 0.001$) correlations to users' qualitative ratings of their experience with live captions. This shows preliminary indications that our proposed flicker metric is an effective metric to quantify the stability of and users' subjective feeling when engaging with live captions.

5.4.2 Stabilization of Live Captions. Crowdworkers had significantly different experiences with the four different conditions (Figure 5). In general, our proposed technique (Stabilized Captions Smooth) received consistently better ratings, significant in five out of six survey questions when compared to the baseline, Raw Google Cloud ASR (Mann-Whitney U test $p < 0.01$) and the confidence thresholding approach, Raw ASR + thresholding (Mann-Whitney U test $p < 0.05$), with a non-significant difference regarding the easiness to follow the video. In addition, all stabilization techniques (conditions 1, 2, 3) were significantly better than the baseline (Mann-Whitney U test $p < 0.05$). Crowdworkers considered the stabilized captions to be more comfortable and easier to read, while feeling less distraction, fatigue, and impairment to their experience than the default live captions and live captions with confidence thresholding. In general, we observed a trend that crowdworkers ranked conditions 3 > 2 > 1 > 0 for all survey questions, with the largest difference observed in distraction, easy to read, fatigue, and impaired viewing experience.

6 LIMITATIONS

Despite the significant findings, several limitations of the present study should be considered. First, the data collection method used in this study did not control for participants' environments or contexts, which may have influenced their subjective ratings. For example, participants may have been in a noisy environment or experienced

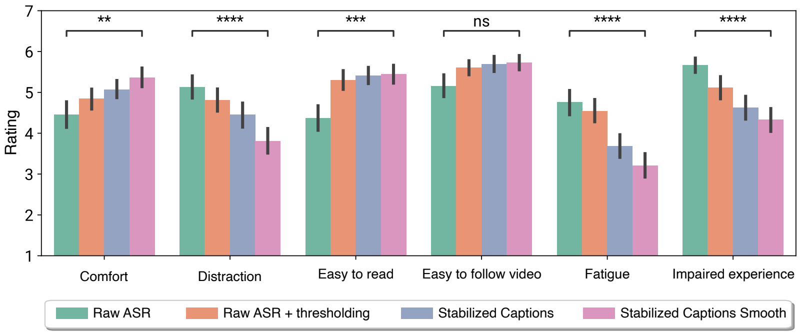


Figure 5: Participants' Task Load Index and Likert scale ratings (from 1 - Strongly Disagree to 7 - Strongly Agree) to four different kinds of live captions: Raw ASR, Raw ASR + thresholding, Stabilized Captions, and Stabilized Captions Smooth, with 95% confidence interval bar. We additionally annotated statistical tests (Mann-Whitney U) results between different conditions: ns (non-significant), * (p<.05), ** (p<.01), * (p<.001), **** (p<.0001).**

interruptions during the caption reading task, which could have affected their perception of the captions. Second, the six questions used in this study were related to the overall reading experience of the captions (e.g., comfort, distraction) rather than specific aspects of the captions, such as accuracy or relevance. While these questions provide valuable insight into the participants' subjective experience, they may not fully capture the extent to which captions improve comprehension. Finally, the study only examined English transcriptions, which may not be representative of other languages, especially non-Latin languages. The cultural and linguistic differences may impact the perception and comprehension of captions and therefore limit the generalizability of the results to other languages and cultures.

7 DISCUSSION AND FUTURE WORK

Language-based Metrics. The flicker metrics provide a low-level measurement of perceptual dispersion when reading live captions. However, they may not fully reflect users' discomfort and comprehension in the wild. In future work, we are interested in purposing language-based metrics that focus on the consistency of the words and phrases used in live captions over time, rather than just the vision-based changes of the captions. These metrics may provide a more accurate reflection of user discomfort and comprehension in real-world scenarios. For example, we could measure the consistency of the words and phrases used in live captions, the frequency of errors made by the ASR system, and changes to words and layouts.

Eye-tracking Lab Study. While our survey provides valuable insights into viewers' experience with live captions, we are also interested in conducting an eye-tracking lab study to better understand how viewers interact with live captions, as well as providing more quantitative evidence of the impact of the stabilization techniques. By tracking viewers' gaze patterns, such as eye fixation and saccades, we could gain a more detailed understanding of the areas of the live captions and the types of errors that are most distracting, and how to improve text stability for those.

Text Stability Beyond Live Captions. We also see a potential for applying the stabilization techniques developed in this work to other applications. For example, the technique could be applied to the field of live translation, where real-time, accurate translations are vital for effective communication. Additionally, the technique could be adapted for use in other contexts, such as live subtitles for movies and television shows, or even in closed captioning for live events such as speeches and performances.

8 CONCLUSION

In this paper, we identify the problem of text stability in live captioning, which can significantly impair users' reading experience. We proposed a quantitative metric to model captions stability using a vision-based flicker metric, and an algorithm to stabilize the rendering of live captions. We conducted a crowdsourced study to evaluate viewers' experience with live captioning. Our results suggest a significant correlation between our proposed flickering metric and viewers' experience, and that our proposed stabilization techniques significantly improved viewers' experience. Our proposed solutions can be integrated into existing ASR systems to enhance the usability of live captions for diverse users, including those with translation needs or those with hearing accessibility needs.

REFERENCES

[1] Katrin Angerbauer, Heike Adel, and Ngoc Thang Vu. 2019. Automatic Compression of Subtitles with Neural Networks and its Effect on User Experience. In *Proc. Interspeech 2019*. Interspeech, New York, NY, USA, 594–598. https://doi.org/10.21437/Interspeech.2019-1750

[2] Jacob Aron. 2011. How innovative is Apple's new voice assistant, Siri?

[3] Larwan Berke, Christopher Caulfield, and Matt Huenerfauth. 2017. Deaf and Hard-of-Hearing Perspectives on Imperfect Automatic Speech Recognition for Captioning One-on-One Meetings. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility* (Baltimore, Maryland, USA) *(ASSETS '17)*. Association for Computing Machinery, New York, NY, USA, 155–164. https://doi.org/10.1145/3132525.3132541

[4] Janine Butler, Brian Trager, and Byron Behm. 2019. Exploration of Automatic Speech Recognition for Deaf and Hard of Hearing Students in Higher Education Classes. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*. Association for Computing Machinery, New York, NY, USA, 32–42. https://doi.org/10.1145/3308561.3353772

[5] Teresa Hirzle, Maurice Cordts, Enrico Rukzio, and Andreas Bulling. 2020. A survey of digital eye strain in gaze-based interactive systems. In *ACM Symposium on Eye Tracking Research and Applications*. Association for Computing Machinery, New York, NY, USA, 1–12.

[6] Ippei Hisaki, Hiroaki Nanjo, and Takehiko Yoshimi. 2010. Evaluation of Speech Balloon Captions for Auditory Information Support in Small Meetings. In *Proceedings of the 20th International Congress on Acoustics*. Association for Computing

Machinery, New York, NY, USA, 1–5. https://doi.org/10.1145/3491102.3501920

[7] Richang Hong, Meng Wang, Mengdi Xu, Shuicheng Yan, and Tat-Seng Chua. 2010. Dynamic Captioning: Video Accessibility Enhancement for Hearing Impairment. In *Proceedings of the 18th ACM International Conference on Multimedia*. Association for Computing Machinery, New York, NY, USA, 421–430. https://doi.org/10.1145/1873951.1874013

[8] Dhruv Jain, Rachel Franz, Leah Findlater, Jackson Cannon, Raja Kushalnagar, and Jon Froehlich. 2018. Towards Accessible Conversations in a Mobile Context for People Who Are Deaf and Hard of Hearing. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility* (Galway, Ireland) *(ASSETS '18)*. Association for Computing Machinery, New York, NY, USA, 81–92. https://doi.org/10.1145/3234695.3236362

[9] Sushant Kafle, Becca Dingman, and Matt Huenerfauth. 2021. Deaf and Hard-of-Hearing Users Evaluating Designs for Highlighting Key Words in Educational Lecture Videos. *ACM Transactions on Accessible Computing (TACCESS)* 14, 4 (2021), 1–24. https://doi.org/10.1145/3470651

[10] Sushant Kafle, Peter Yeung, and Matt Huenerfauth. 2019. Evaluating the Benefit of Highlighting Key Words in Captions for People Who Are Deaf or Hard of Hearing. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility* (Pittsburgh, PA, USA) *(ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 43–55. https://doi.org/10.1145/3308561.3353781

[11] Saba Kawas, George Karalis, Tzu Wen, and Richard E Ladner. 2016. Improving Real-Time Captioning Experiences for Deaf and Hard of Hearing Students. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility*. Association for Computing Machinery, New York, NY, USA, 15–23. https://doi.org/10.1145/2982142.2982164

[12] Jan-Louis Kruger, Esté Hefer-Jordaan, and Gordon Matthew. 2013. Measuring the Impact of Subtitles on Cognitive Load: Eye Tracking and Dynamic Audiovisual Texts, In ACM. *ACM International Conference Proceeding Series* 1, 62–66. https://doi.org/10.1145/2509315.2509331

[13] Jan-Louis Kruger and Faans Steyn. 2013. Subtitles and Eye Tracking: Reading and Performance. *Reading Research Quarterly* 49 (10 2013). https://doi.org/10.1002/rrq.59

[14] Kuno Kurzhals, Emine Cetinkaya, Yongtao Hu, Wenping Wang, and Daniel Weiskopf. 2017. Close to the Action: Eye-Tracking Evaluation of Speaker-Following Subtitles. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 6559–6568. https://doi.org/10.1145/3025453.3025772

[15] Raja S Kushalnagar, Walter S Lasecki, and Jeffrey P Bigham. 2012. A Readability Evaluation of Real-Time Crowd Captions in the Classroom. In *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility*. Association for Computing Machinery, New York, NY, USA, 71–78. https://doi.org/10.1145/2384916.2384930

[16] Raja S Kushalnagar, Walter S Lasecki, and Jeffrey P Bigham. 2014. Accessibility Evaluation of Classroom Captions. *ACM Transactions on Accessible Computing (TACCESS)* 5, 3 (2014), 1–24. https://doi.org/10.1145/2982142.2982164

[17] Seongjae Lee, Sunmee Kang, Hanseok Ko, Jongseong Yoon, and Minseok Keum. 2013. Dialogue Enabling Speech-to-Text User Assistive Agent With Auditory Perceptual Beamforming for Hearing-Impaired. In *2013 IEEE International Conference on Consumer Electronics (ICCE)*. IEEE, IEEE, New York, NY, USA, 360–361. https://doi.org/10.1109/ICCE.2013.6486929

[18] Kirill Levin, Irina Ponomareva, Anna Bulusheva, German Chernykh, Ivan Medennikov, N. Merkin, and Natalia Tomashenko. 2014. Automated Closed Captioning for Russian Live Broadcasting. In *Interspeech*. Interspeech, New York, NY, USA. https://doi.org/10.21437/Interspeech.2014-352

[19] Zehan Li, Haoran Miao, Keqi Deng, Gaofeng Cheng, Sanli Tian, Ta Li, and Yonghong Yan. 2022. Improving Streaming End-to-End ASR on Transformer-Based Causal Models With Encoder States Revision Strategies. https://doi.org/10.48550/ARXIV.2207.02495

[20] Xingyu Liu, Vladimir Kirilyuk, Xiuxiu Yuan, Peggy Chi, Xiang Chen, Alex Olwal, 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)*. Association for Computing Machinery, New York, NY, USA, 14. https://doi.org/10.1145/3544548.3581566

[21] Xingyu "Bruce" Liu, Ruolin Wang, Dingzeyu Li, Xiang "Anthony" Chen, and Amy Pavel. 2022. CrossA11y: Identifying Video Accessibility Issues via Cross-Modal Grounding. 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 43, 14 pages. https://doi.org/10.1145/3526113.3545703

[22] Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. 2018. STACL: Simultaneous Translation with Implicit Anticipation and Controllable Latency using Prefix-to-Prefix Framework. In *Annual Meeting of the Association for Computational Linguistics*. Association for Computing Machinery, New York, NY, USA, 10.

[23] Tara Matthews, Scott Carter, Carol Pai, Janette Fong, and Jennifer Mankoff. 2006. Scribe4Me: Evaluating a Mobile Sound Transcription Tool for the Deaf. In *International Conference on Ubiquitous Computing*. Springer, Springer, New York, NY, USA, 159–176. https://doi.org/10.1007/1185356_10

[24] Mohammad Reza Mirzaei, Seyed Ghorshi, and Mohammad Mortazavi. 2012. Using Augmented Reality and Automatic Speech Recognition Techniques to Help Deaf and Hard of Hearing People. In *Proceedings of the 2012 Virtual Reality International Conference*. Association for Computing Machinery, New York, NY, USA, 1–4. https://doi.org/10.1145/2331714.2331720

[25] Mark Neerincx, Anita Cremers, Judith Kessens, David Van Leeuwen, and Khiet Truong. 2009. Attuning Speech-Enabled Interfaces to User and Context for Inclusive Design: Technology, Methodology and Practice. *Universal Access in the Information Society* 8 (06 2009), 109–122. https://doi.org/10.1007/s10209-008-0136-x

[26] Thai Son Nguyen, Jan Niehues, Eunah Cho, Thanh-Le Ha, Kevin Kilgour, Markus Muller, Matthias Sperber, Sebastian Stueker, and Alex Waibel. 2020. Low Latency ASR for Simultaneous Speech Translation. https://doi.org/10.48550/ARXIV.2003.09891

[27] Alex Olwal, Kevin Balke, Dmitrii Votintcev, Thad Starner, Paula Conn, Bonnie Chinh, and Benoit Corda. 2020. Wearable Subtitles: Augmenting Spoken Communication With Lightweight Eyewear for All-Day Captioning. In *UIST '20: Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, NY, USA, 10. https://doi.org/10.1145/3379337.3415817

[28] Agnès Piquard-Kipffer, Odile Mella, Jérémy Miranda, Denis Jouvet, and Luiza Orosanu. 2015. Qualitative Investigation of the Display of Speech Recognition Results for Communication With Deaf People. In *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*. Association for Computing Machinery, New York, NY, USA, 36–41.

[29] Soraia Silva Prietch, Napoliana Silva de Souza, and Lucia Villela Leite Filgueiras. 2014. A Speech-to-Text System's Acceptance Evaluation: Would Deaf Individuals Adopt This Technology in Their Lives?. In *International Conference on Universal Access in Human-Computer Interaction*. Springer, Springer, New York, NY, USA, 440–449. https://doi.org/10.1145/3290607.3312921

[30] Nils Reimers and Iryna Gurevych. 2019. Sentence-Bert: Sentence Embeddings Using Siamese Bert-Networks. *ArXiv Preprint ArXiv:1908.10084* 1 (2019), 10. https://arxiv.org/pdf/1908.10084

[31] Pablo Romero-Fresco. 2015. *Accuracy Rate in Live Subtitling: The NER Model.* Palgrave Macmillan UK, London, 28–50. https://doi.org/10.1057/978113755289_3

[32] Yuan Shangguan, Kate Knister, Yanzhang He, Ian McGraw, and Françoise Beaufays. 2020. Analyzing the Quality and Stability of a Streaming End-to-End On-Device Speech Recognizer. *CoRR* abs/2006.01416 (2020), 10. arXiv:2006.01416 https://arxiv.org/abs/2006.01416

[33] Brent N Shiver and Rosalee J Wolfe. 2015. Evaluating Alternatives for Better Deaf Accessibility to Selected Web-Based Multimedia. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*. Association for Computing Machinery, New York, NY, USA, 231–238. https://doi.org/10.1145/2700648.2809857

[34] ACM SIGCHI. 2. *CHI 2022 Town Hall.* ACM SIGCHI. https://www.youtube.com/watch?v=dDPPNyUDmco

[35] Michael S Stinson, Lisa B Elliot, and Ronald R Kelly. 2017. Deaf and Hard-of-Hearing High School and College Students' Perceptions of Speech-to-Text and Interpreting/note Taking Services and Motivation. *Journal of Developmental and Physical Disabilities* 29, 1 (2017), 131–152. https://doi.org/10.1177/0022466907313453

[36] Michael S Stinson, Lisa B Elliot, Ronald R Kelly, and Yufang Liu. 2009. Deaf and Hard-of-Hearing Students' Memory of Lectures With Speech-to-Text and Interpreting/note Taking Services. *The Journal of Special Education* 43, 1 (2009), 52–64.

[37] Prakhar Swarup, Roland Maas, Sri Garimella, Sri Harish Mallidi, and Björn Hoffmeister. 2019. Improving ASR Confidence Scores for Alexa Using Acoustic and Hypothesis Embeddings.. In *Interspeech*. Interspeech, New York, NY, USA, 2175–2179.

[38] Stefan Winkler, Elisa Drelie Gelasca, and Touradj Ebrahimi. 2003. Toward Perceptual Metrics for Video Watermark Evaluation. In *Applications of Digital Image Processing XXVI*, Andrew G. Tescher (Ed.), Vol. 5203. International Society for Optics and Photonics, SPIE, New York, NY, USA, 371 – 378. https://doi.org/10.1117/12.512550

[39] Yuekun Yao and Barry Haddow. 2020. Dynamic Masking for Improved Stability in Online Spoken Language Translation. In *Conference of the Association for Machine Translation in the Americas*. Association for Computing Machinery, New York, NY, USA, 10.

A STABILIZED CAPTIONS NORMALIZATION FOR FLICKER METRICS

Following Winkler *et al.* [38], we divide flicker metrics by a normalization factor *m* for each frame by taking maximum possible values

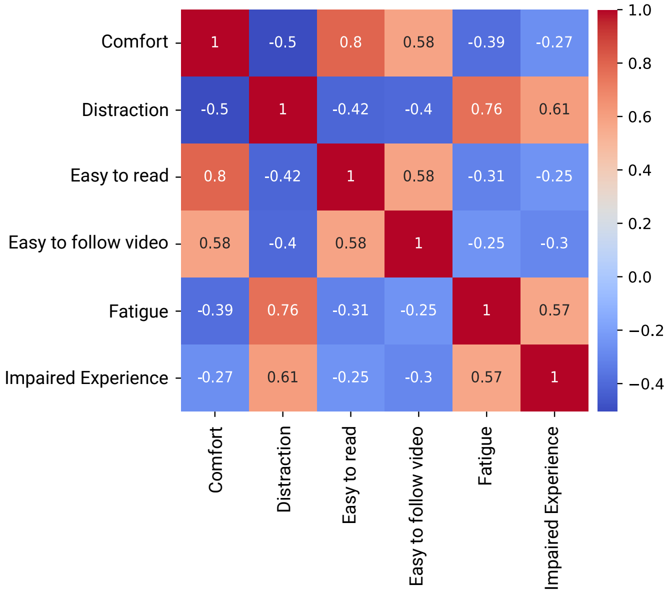


Figure 6: Correlation matrix of users' experience with live captions (Likert scale ratings) in our user study. In particular, we noticed that the "comfort" and "easy to read captions" are highly correlated; "distraction" and "eye fatigue" are highly correlated.

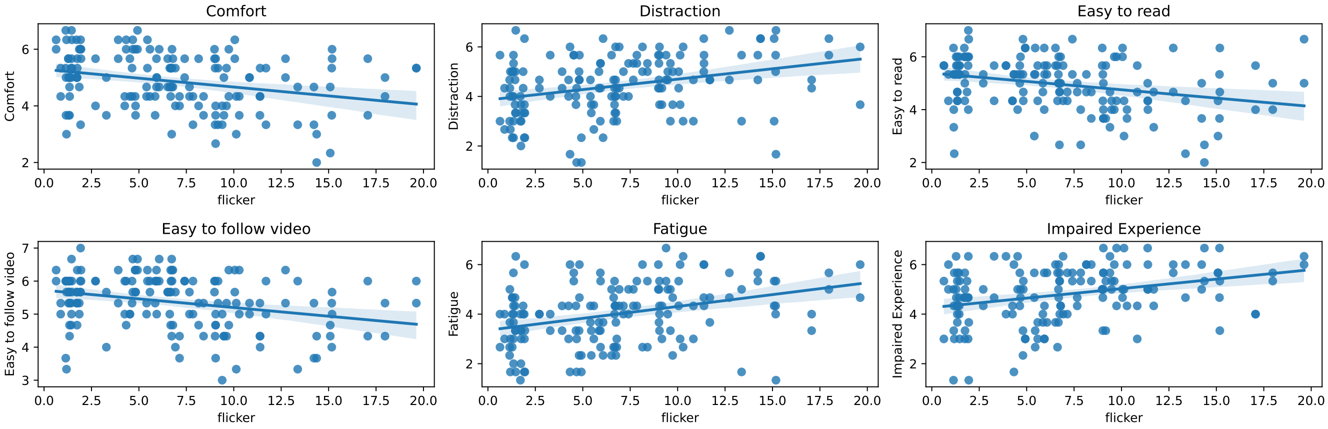


Figure 7: Scatter plots of our purposed flicker metric vs. users' Likert scale ratings in our user study.

of both spatial and temporal domains into account. We define S_i as the gradients computed from the horizontally and vertically Sobel-filtered image $\mathbf{I}_i$, and $\mathbf{D}_i$ as the temporal difference. We then set the normalization factor as the product of the $m = \max\left(\overline{S_i} \cdot \overline{\mathbf{D}_i}, \delta\right)$. $\delta = 0.007$ is a threshold to avoid extreme values of normalization. Using and without using this normalization factor does *not* yield much difference in the final outcome for our data, except for scaling the values. We provided an implementation of this metric in the supplementary material.

B STABILIZED CAPTIONS DATA ANALYSIS DETAILS

We present data analysis details on users' perception of caption stability. We examined the correlation matrix (Figure 6) between users' experience with live captions, which were rated on a Likert scale. Our analysis revealed a strong correlation between the ratings for "comfort" and "ease of reading captions", as well as a significant correlation between "distraction" and "eye fatigue". In addition, we present scatter plots (Figure 7) comparing our proposed flicker metric with users' Likert scale ratings.