Requirements Engineering (2023) 28:521–539
https://doi.org/10.1007/s00766-023-00403-0

ORIGINAL ARTICLE

When details are difficult to portray: enriching vision videos

Lukas Nagel[1] · Melanie Schmedes[1] · Maike Ahrens[1] · Kurt Schneider[1]

Received: 17 January 2023 / Accepted: 22 June 2023 / Published online: 5 September 2023

Abstract
The creation of a shared understanding of the project vision of all relevant stakeholders is vital to the requirements engineering process. One way to create such a shared understanding is through the use of vision videos that visualize the project vision at an early project stage. However, not all functional aspects can be presented. For example, the fact that an access code is valid for only a single use can be hard to visualize. One low-effort solution could be the insertion of short texts or short audio clips. In this work, our question is twofold: What effects do short pieces of additional information have in vision videos? What are suitable ways to add this information to vision videos? To answer these research questions, we investigated three different methods of inserting additional information to vision videos in an eye tracking study. We inserted short texts either below the scene or as overlays and also investigated the addition of short audio clips. These methods were evaluated in terms of participants' video comprehension, visual effort, cognitive load and subjective preference. The results of our study show that the pieces of additional information improve vision comprehension, thereby supporting the creation of a shared understanding. All investigated methods lead to only marginal increases of the viewers' cognitive load. Based on our results, we derive recommendations on how to insert additional information in vision videos.

Keywords Requirements engineering · Eye tracking · Vision · Video

1 Introduction

One factor impacting the success of software projects is the degree of shared understanding among all relevant stakeholders regarding the future system [1, 2]. Creating such a shared understanding can be especially difficult as different people have different mental models of themselves and their environment [3, p.17]. Requirements engineers face the challenge of aligning these mental models to form a common vision. One possible solution is using videos in addition to written documents. *Vision Videos* present the future system in a medium with high communication richness and effectiveness [4, 5] and can serve as a starting point to foster meaningful discussions among stakeholders [6]. These discussions can lead to the uncovering and subsequent resolving of misunderstandings, thereby achieving a shared vision [6].

Vision videos can be used at different stages of the software development process [7]. One example is the application at a very early project stage to communicate and validate the project vision [7]. In such instances, video creators might not have access to a prototype of the system. Additionally, the creation of vision videos also has to adhere to time and resource constraints in order to fit into the development plans. These differences in the available resources between vision videos and high-end video productions have lead to the introduction of the *Affordable Video Approach* by Schneider et al. [8]. According to this approach, individual parts of the project vision that might not have been implemented in a prototype can be substituted by actors simply pretending to use the system as envisioned. For example, a newly conceptualized remote control that enables parcel carriers to access the trunks of recipients' cars could be substituted by an actor using a common TV remote.

✉ Lukas Nagel
lukas.nagel@inf.uni-hannover.de

Melanie Schmedes
melanie.schmedes@inf.uni-hannover.de

Maike Ahrens
maike.ahrens@inf.uni-hannover.de

Kurt Schneider
kurt.schneider@inf.uni-hannover.de

[1] Software Engineering Group, Leibniz University Hannover, Welfengarten 1, 30167 Hannover, Lower Saxony, Germany

However, not all details of the envisioned functionality can be substituted in this way. Details of the aforementioned example, like the fact that the remote access functionality exclusively opens the trunk and only a single time, can be challenging to visualize in video content. It is also not feasible to dedicate the required screen time to portray each detail in its entirety. Such an approach would lengthen the vision video beyond the recommended maximum length of 5 min [9] and therefore defy its inherent purpose. In longer videos stakeholders might miss important misalignments between the system vision presented in the video and their own. A further issue can be found when the need for a clarification of specific details is discovered only after the video has been produced. In all of these cases, the insertion of additional information to existing video content can provide the required details without requiring any reshooting or reproduction processes.

Affordable ways for the insertion of additional information to existing vision videos are needed to keep the effort for requirements engineers at a moderate level. One possible low-effort solution could be the enrichment of the video content with short texts. Short pieces of narrated audio might also present a suitable alternative that is even less invasive to the visual content. We examine these potential solutions in an eye tracking study. Our goal is to *find a suitable method of inserting additional information to vision videos*.

In this paper, we present three different designs for the insertion of additional information to existing vision videos. These designs include a variant to add short texts below the video scene, one variant with short texts as overlays and a variant to add a narrator voice providing the same information via audio. The results of our eye tracking study show that the participants were able to reproduce about two thirds of the additional information after watching the video for the first time. Self-reported cognitive load measures and visual effort metrics obtained with the eye tracker did not present statistically significant drawbacks of the variants. The methods evaluated in this paper present suitable ways of inserting additional information to existing vision videos. We combine our eye tracking data with subjective feedback obtained in a questionnaire to investigate the designs in terms of video comprehension, reading time, cognitive load and subjective preference.

The presented paper is an extension of the previous work of Schmedes et al. [10]. It complements it by evaluating further opportunities to convey information in already existing videos. Instead of solely looking at text options, we also included a variant with additional audio clips in which the information is provided by a narrator. We evaluated the advantages and drawbacks of this option by extending the conducted eye tracking study. Besides, we expanded the statistical analysis and discussion of the existing results. Moreover, we included additional related work, e.g., of using audio in contexts of multimedia learning.

The paper is structured as follows: Sect. 2 provides information on the background of our research and related work. The eye tracking study is described in Sect. 3 and its results are lined out in Sect. 4. These results are discussed in Sect. 5 before Sect. 6 concludes the paper.

2 Background & related work

The topics of this work have already been researched and discussed in other scientific literature which presents the background of our work.

2.1 Videos in requirements engineering

The use of videos in requirements engineering has been examined in multiple related works. One of the earliest attempts of the use of video in the field of software engineering was an approach by DeMarco and Geertgens who presented a program documentation on VHS tapes [11]. Later, so-called vision videos emerged who can be used to visualize concrete scenarios [12, 13] or product visions [7, 14]. Differences of vision videos and use cases were evaluated in a work by Brill et al. [7]. Their work focused on vision videos that were created quickly and with little effort. A refinement of a vision through the presentation of alternatives in videos was explored by Schneider et al. [8]. Busch et al. [15] compared animated vision videos with those created with actors. They concluded that animated videos appear to be an adequate alternative.

Guidelines supporting the production of vision videos have been summarized by Karras and Schneider [9]. They provide detailed support for the phases of pre-production, shooting and postproduction. A further work by Karras et al. [4] presents a quality model for vision videos. Results of their experiment lead to the identification of six characteristics vital to a video with high quality, namely *video length, focus, prior knowledge, clarity, pleasure* and *stability*.

Another aspect of research interest regarding vision videos is how they should be used [16, 17]. One possible usage scenario that has been examined by Karras et al. [16] is the context of a virtual focus group. Nagel et al. [17] studied asynchronous and synchronous viewing contexts for vision videos. They conducted an online study and formulated recommendations on which context is applicable to which situation.

2.2 Software engineering, requirements engineering and eye tracking

A systematic literature review on eye tracking in software engineering has been performed by Sharafi et al. [18]. Their

paper summarizes the usage of eye tracking and the measurements taken in previous studies. They also point out limitations of eye tracking. Additionally, their systematic literature review identified the main topics of research papers on eye tracking in software engineering including *model comprehension*, *code comprehension*, *debugging*, *collaborative instructions* and *traceability* [18].

In the field of requirements engineering, Ahrens et al. [19] investigated how specifications are read in an eye tracking study. They also compared specifications on paper and on screen and found reading patterns that appeared to be similar. Ahrens and Schneider [20] studied three different attention representation types to support the reading of specifications. These types were based on previously gathered eye tracking data. The considered representation types were quick access buttons, heatmap bars and role icons [20].

Eye tracking has also been used to determine how use cases and linked requirements are read. Karras et al. [21] compared three different linking variants. Use cases were read before requirements in all cases. In a follow-up eye tracking experiment, Karras et al. [22] studied the reading behavior of use cases and requirements to corroborate their previous results [21].

Gralha et al. [23] examined a number of user story templates regarding the aspects of user story creation and understanding. One of their metrics was the visual effort experienced by their participants which they measured using an eye tracker. Busjahn et al. [24] looked to identify a starting point for understanding software code by investigating how code is read with an eye tracker. For example, they compared the reading of code and natural language texts. A research effort by Jermann and Nüssli [25] investigated gaze behavior and cooperation during pair programming with eye tracking. They report that pairs mostly looked at the same screen elements during spoken selections.

2.3 Eye tracking of videos

A study by Goldstein et al. [26] investigated whether people look at the same screen elements while watching movies. They created COIs (Center of Interest) based on eye tracking data and concluded that these COIs could be used for magnification to help visually impaired viewers [26]. Srivastava et al. [27] studied the connection between paying attention to the video content and listening to a voice in the context of educational videos. They found a relation between the gaze behavior of their participants and prior knowledge [27].

A paper by Brown et al. [28] details their investigation of differences between common subtitles placed below the video and dynamic subtitles superimposed as overlays. Their eye tracking study revealed that most participants had a positive attitude toward dynamic subtitles. One example for an advantage of these dynamic subtitles was that they made it easier to focus on the content of the video. Nevertheless, Brown et al. [28] found that no general preference of dynamic subtitles was evident in their results. Therefore, they recommend that viewers should be able to choose a subtitle design.

Kruger et al. [29] conducted an eye tracking study to investigate the impact of subtitles on the cognitive load of viewers. They also recorded Electroencephalography-measurements, a comprehension test and a questionnaire. No significant differences in terms of short- or long-term performance measures could be found between the groups with and without subtitles.

2.4 Eye tracking and text

Related research has addressed various subtopics of the manner in which text is read by humans. Rayner and McConkie [30] investigated different models of eye guidance in reading and conducted a study regarding eye movements. Their conclusions include that the fixation duration at least partially depends on features of the text that is being fixated [30]. Dogusoy et al. [31] found that sans-serif fonts are read faster and more precisely. A further eye tracking study by Rayner [32] investigated reading behaviors. The study's results indicate that eye movements are influenced by cognitive processes that take place at the fixation moment. Research performed by Hall and Hanna [33] examined the color of text and backgrounds of web pages and their effect on one another. Aspects addressed in their research include *readability*, *retention*, *aesthetics* and *behavioral intentions*. They also found that colors with a higher ratio of contrast result in better readability [33].

2.5 Visual and auditory information

Mayer [34] sums up and explains the theoretical basis for a cognitive theory of multimedia learning. In this context, Mayer [34] explains the assumptions of dual channels, limited capacity and active processing. The term dual-channel assumption describes that humans have two different channels of perception, one for visual and one for auditory information [35].

Both kinds of information are in a first step stored briefly in a visual respectively auditory register [36]. Subsequently, the processing takes place in the working memory, whose capacity is limited [36]. Visual working memory includes an average of five units and auditory includes articulations in the time frame of about two seconds [36].

Simpson and Thomas [37] focused on different presentation modes to study listening and reading comprehension. Within the scope of a study with 368 students, they compared listening to a lecture, listening to a tape recording of a lecture, reading a text about the content and structure

of the lecture, and reading a text in which the ideas were emphasized by underlining and capitalization [37]. After the subjects studied the specific material, a content test took place. The results of the groups did not differ. Simpson and Thomas [37] formulated that their findings would support the assumption that listening and reading comprehension are based on the same underlying processes. Results of the later test show a better understanding of the audio group. According to Simpson and Thomas [37] this result would strengthen the theoretical point of view that spoken and written language have their own and/or singular processes.

According to the guidelines for combining media of the DIN EN ISO 14915-3:2002 [38] software ergonomics for multimedia user interfaces, speech has the ability to strengthen information in movies. Moreover, speech can draw attention to specific film elements. In addition, the use of speech is described as an alternative option to text, which may complicate viewing the image in some circumstances.

Richards [39] formulated that three levels seem to be important for listening: (1) identification of statements, (2) interpretation of illocutionary power, and (3) mobilization of knowledge about the real world. The long-term memory is dealing with concepts and meanings instead of the form or distinct sentences [39].

2.6 Scope of this work

The research field of vision videos has been explored by multiple research efforts. These efforts include the affordable video approach [8, 15] or the development of guidelines [9]. Research in the field of eye tracking has examined many different aspects related to requirements engineering, texts and videos. The preceding sections of related work deal with thematic sub-aspects which we looked to combine for the context of this work. We looked to use an eye tracker and a questionnaire to investigate the effects of inserting additional information via text or via audio. The texts we used differed from regular subtitles as we looked to integrate short pieces of additional information rather than transcribing the video content or spoken dialog. We thereby focused on the opportunity of enriching and adapting existing vision videos with additional information while minimizing the effort of making these adaptations.

3 Eye tracking study

We performed an eye tracking study to reach our goal of *finding a suitable method of inserting additional information to vision videos*. To ensure that we could find an appropriate experiment design, we applied the GQM paradigm [40]. Therefore, we defined our research goal before determining fitting research questions and metrics. The definition of our research goal was performed using Wohlin et al.'s [41] goal definition template.

Goal definition:
We analyze three methods of adding information to vision videos
for the purpose of assessing information retention
with respect to the comprehension of the video content
from the point of view of requirements engineers
in the context of an experimental setting simulating early RE

To reach this goal, we designed a between-groups eye tracking study. A within-group study design was not suited to our research goals, as we would have to either accept the impact of a strong learning effect on our data or find two vision videos that were of similar complexity. We decided that a learning effect was unacceptable. Finding two vision videos of similar complexity was impossible as the number of criteria that could potentially impact the complexity perceived by individual participants. We also preferred to accept the threat of a low sample size, which could be solved with later replications of the study, over the issues stemming from a crossover design [42].

We tested three different methods of adding information to already existing vision videos. The first two methods inserted additional information by adding text. The variant *TextBelow* presented the text below the screen space of the original video, akin to how subtitles are commonly shown. The variant *TextOverlay* displayed the same text as a block superimposed on the video content in a treatment-specific place. Lastly, the *AudioVariant* conveyed the information by adding a narrative voice to the existing audio track. We also conducted sessions with a control group that watched the original vision video without any added information. Figure 1 shows an example of the four different variants.

The use of an eye tracker enabled a detailed determination of whether the textual additions are recognized and read by the viewer. Eye tracking also enables the measurement of metrics for the visual effort of participants [43, 44] which provides a basis for a more detailed evaluation of the three methods.

3.1 Research questions & hypotheses

Our research goal is refined by the following research questions. We abbreviate the name of the group watching the *TextBelow* variant as *TB*. The abbreviations *TO* for the *TextOverlay* variant, *AV* for the *AudioVariant* and *CG* for the control group variant are used accordingly.

Research Question 1:
In what ways is the viewing behavior influenced by additional information when watching vision videos?

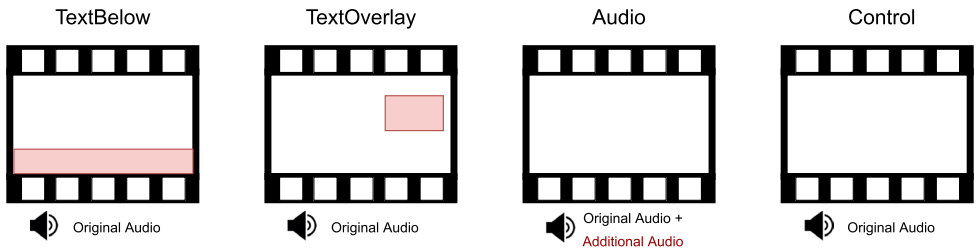


Fig. 1 An overview of the different video variants used in the study

First, the impact of the different variants on the viewing behavior of our participants is determined. We therefore analyze the visual effort evident in the eye tracking data. We developed hypothesis 1 and the corresponding subhypotheses to refine Research Question 1. We also tested the alternative hypotheses:

H1$_0$: There is no difference between the groups with additional information in viewing the videos in terms of visual effort.

H1$_{i.0}$: There is no difference between group a and group b in viewing the videos in terms of visual effort.

$$i = (a, b) \; with \; a \neq b,$$
$$a, b \in \{TB, TO, AV\}$$

Research Question 2:
What type of information presentation is most suitable for the context of vision videos?

The second research question aims to establish the suitability of the three methods of adding information. We consider the perceived cognitive load and vision comprehension. Research Question 2 is refined through two hypotheses and subhypotheses. We also tested the corresponding alternative hypotheses.

H2$_0$: There is no difference between the four groups regarding the perceived cognitive load.

H2$_{j.0}$: There is no difference between group a and group b regarding the perceived cognitive load.

$$j = (a, b) \; with \; a \neq b,$$
$$a, b \in \{TB, TO, AV, CG\}$$

H3$_0$: There is no difference between the three groups with additional information in terms of the amount of newly gathered knowledge on the content of the video.

3.2 Metrics

To answer our research questions, various metrics were determined with respect to the hypotheses. All measurements were obtained from the eye tracking data using Tobii Pro Lab[1] or the questionnaire. Table 1 provides an overview of all metrics. The independent variables of our study are the four variants that were tested, namely the *TextBelow*, *TextOverlay*, *Audio* and *Control* variants.

The visual effort experienced while watching the different vision video variants was investigated in terms of the average fixation duration and the fixation and saccade count. Fixations describe a stable eye gaze that usually lasts for 100 to 300ms. Saccades are the rapid eye movements in-between. For the treatment groups watching one of the two *TextVariants*, we also examined the overall fixation time spent looking at the text as well as the visit count, meaning how often the text was looked at. We picked these metrics based on work by Sharafi et al. [43], who reported that more fixations on an area indicate more visual attention. Additionally, work by Jeanmart et al. [44] laid out that a higher complexity or larger importance of viewing areas can be indicated by longer fixations or a longer overall fixation time. As for the Likert scales selected for the video and text characteristics, we decided to use six-point Likert scales to avoid a neutral position. This was suited to our evaluation as we expected viewers to have an opinion leaning to one side (e.g., regarding the question "Did the videos help to understand the visions?"). The lack of a neutral option enabled more powerful results to be obtained. The question asking participants about their cognitive load was designed according to the nine-point Paas scale [45]. This specialized Likert scale ranges from values of 1 indicating a very very low mental effort like the one experienced while riding a bike to a value of 9 which denotes a very very high mental effort similar to what might be experienced while taking an examination. It has been proven to be a reliable measure for the cognitive load of users of a system [46].

[1] https://www.tobiipro.com/product-listing/tobii-pro-lab/.

Table 1 Overview of applied metrics

Dependent variable	Eye tracking metrics
Viewing behavior	Reading time of texts (s), Number of gaze visits
Visual effort	Average fixation duration (ms), Fixation count, Fixation time (ms, Saccade count, Visit count of text elements
Dependent variable	**Metrics observed in questionnaire**
Perceived cognitive load	Rating on a nine-point Likert scale (Paas scale [45])
Correct answers	Number of correctly answered questions about the additional information
Video characteristics	Rating on a six-point Likert scale
Text characteristics	Rating on a six-point Likert scale

3.3 Material

All video variants watched by participants of our study are adaptations of the same vision video.[2] The video portrays different methods for ordering and delivery processes in rural areas. The video's protagonist is shown ordering a product by focusing on it with their smartphone's camera and ordering via a simple press of a button that is associated with the product. In the third ordering vision, the video shows the product's container automatically detecting a low fill-level and placing an order without any direct user interaction. As for the delivery methods, the video presents another set of three options. First, a method is portrayed in which a neighbor was able to collect the package in the city. The second shown method consists of a delivery by drone, while the last method presents a parcel delivery worker leaving the package in the trunk of the recipient's car. The first part of the video presenting the ordering methods runs for 1:22 min, while the second part on the delivery options is 2:45 min long. In total, participants of our study spent 4:07 min watching video content. The vision video was produced in the context of the paper *Refining Vision Videos* [8] and has already been used in multiple other studies [8, 15, 17]. The videos were watched on 24-inch screens equipped with Tobii X3-120 eye trackers.

We inserted additional information to each option. This information could not be gathered from the original vision video. We derived these pieces of information from questions posed by participants of an earlier study which focused on synchronous and asynchronous viewing of vision videos [17]. The information was either presented in short texts or as additional audio voiced by a narrator. The texts ranged from a length of 7 to 10 words (50 to 60 characters). Each text was displayed for a duration of 3 s, meaning that the treatment groups watching the two *TextVariants* were presented with texts for a total of 18 s of runtime. For the *AudioVariant*, a narrator read the same pieces of information out loud. To keep in line with the *TextVariants*, the narrator kept their voice recordings to a similar length of about 3 s.

Overall, four different variants of the vision video were used in the study. One group of participants was shown the *TextBelow* variant which displayed the text below the video, while a second group was shown *TextOverlay* with the text superimposed on the video. A third group watched the original video content enhanced with the audio track of the narrator in the *AudioVariant*. The fourth and final group watched the vision video in its original version as a control group.

A second part of the study included questionnaires[3] that were handed out to participants on paper. Participants were asked for demographic data and to give short summaries of the ordering and delivery methods they had seen in the video. Additionally, the questionnaires included questions about participants' opinions on the content of the videos and questions specific to the additional information that was presented to the treatment groups. The treatment groups were also asked questions regarding the method of adding information that they had encountered. In this way, we were able to collect data on specific characteristics of the text and audio elements. These included the appearance of the individual elements and participants' ability to reproduce the portrayed information.

Table 2 presents an excerpt of different parts of the questionnaire, the related response types and which groups were asked. A subset of the included questions has already been used in prior work [8, 15]. The treatment groups were asked comprehension questions regarding the content of the additional information. An example of such a question can be found in the first question presented in Table 2. Row 3 of the same table gives an example for a question referring to participants' preferences regarding the method in which

[2] The videos used cannot be shared for privacy reasons.

[3] Unfortunately, we cannot share the raw data as participants did not consent to a sharing of their individual data on external servers.

Table 2 Excerpt of the questionnaire. The full set is available on Zenodo [47]

Group	Questions / Statements	Type of answer
Treatment Groups	Method A: How long do you need to focus the product with the camera to place an order?	Free-text field
Treatment Groups	Method 3: How does the access to the recipient's trunk work?	Free-text field
Treatment Groups	Did the texts/audio help to understand the visions?	Likert scale [0-5]
Treatment Groups Control Group	How much mental effort did you invest while watching the vision videos?	Likert scale [0-5]
Treatment Groups Control Group	I liked the videos.	Likert scale [0-5]
Treatment Groups Control Group	The video quality was sufficient to understand the content.	Likert scale [0-5]

information can be added to vision videos. Furthermore, participants of all groups were asked to rate the mental effort they experienced while watching the vision video and to give their opinions on the video in its entirety, like in the final two questions of Table 2.

3.4 Participants

A total of 32 participants took part in the study. An initial set of 24 participants was randomly and evenly distributed among the control group and the two treatment groups watching the *TextVariants*. The final 8 participants were included in a follow-up study in which we examined the *AudioVariant*. This resulted in four different groups of eight participants each. To find our participants we used an external platform connecting experimenters and university students who are looking to participate in scientific studies. The platform allowed for an email to be sent out to potential participants who were then able to select a time slot at which they wanted to take part in the study. These time slots were then assigned to the different groups to ensure that we obtained the same sample size for each treatment. By using the external platform, we had no influence on which of the contacted participants took part and which time slot they selected.

Our participants ranged from 19 to 29 years of age (M = 23.1, SD = 4.7). 15 participants identified themselves as female and 17 as male. 29 of them indicated that they were university students at the time of their participation, 4 indicated themselves to be employed. A single participant stated that they were currently fulfilling both roles. Overall, the 32 participants indicated 19 different fields of study or employment.

Another part of our demographics questionnaire asked how often the participants had been involved in software development processes. For this question we made use of a Likert scale ranging from 0 (never before) to 5 (very often). 19 participants indicated that they had never been involved in software development. The remaining 13 answers ranged from 1 to 3.

Participants received a monetary incentive of 10 euros for their participation in the study. However, this incentive was in no way connected to their performance in the study.

3.5 Experiment setting

Participants took part in individual sessions. These sessions were conducted in a quiet room and adhering to a strict hygiene concept due to the ongoing COVID-19 pandemic. This included only a single experimenter being present during the study as well as regular use of disinfectants on all surfaces that participants might have come in contact with.

Complying with this hygiene concept and social distancing regulations, an overview of the experiment and a consent form were handed out. Next, the process of the study was explained by the experimenter and a brief introduction of the context of the vision video was given. This context was the ordering and delivery processes in rural areas where local stores might be out of reach in modern times due to previous store closings and the urban sprawl.

After these introductory remarks, the eye tracker was calibrated using a 5-point calibration built in Tobii Pro Lab. As a next step, participants were shown the vision video. As determined by our choice of a between-groups design, each participant watched only a single variant of the vision video. They saw one of the following variants: (1) The *TextBelow* variant which presented the additional information below the video, (2) The *TextOverlay* variant showing the text superimposed on the video content, (3) The *AudioVariant* that included an additional audio track voiced by a narrator, or (4) the *ControlVariant* with a version of the vision video without any additional information if they were part of the control group.

Once the respective video had been watched, questionnaires were handed out as the final part of the study. These questionnaires included all questions regarding the video

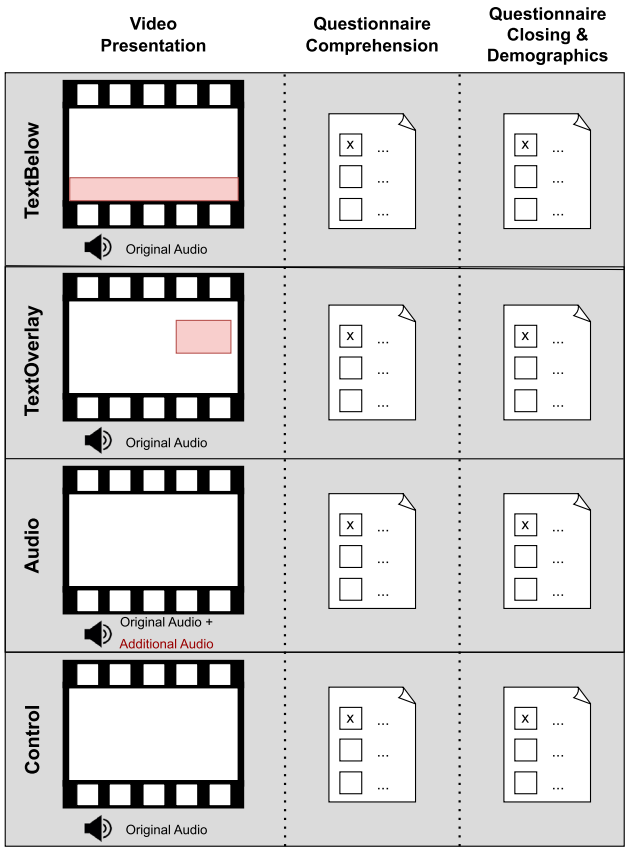


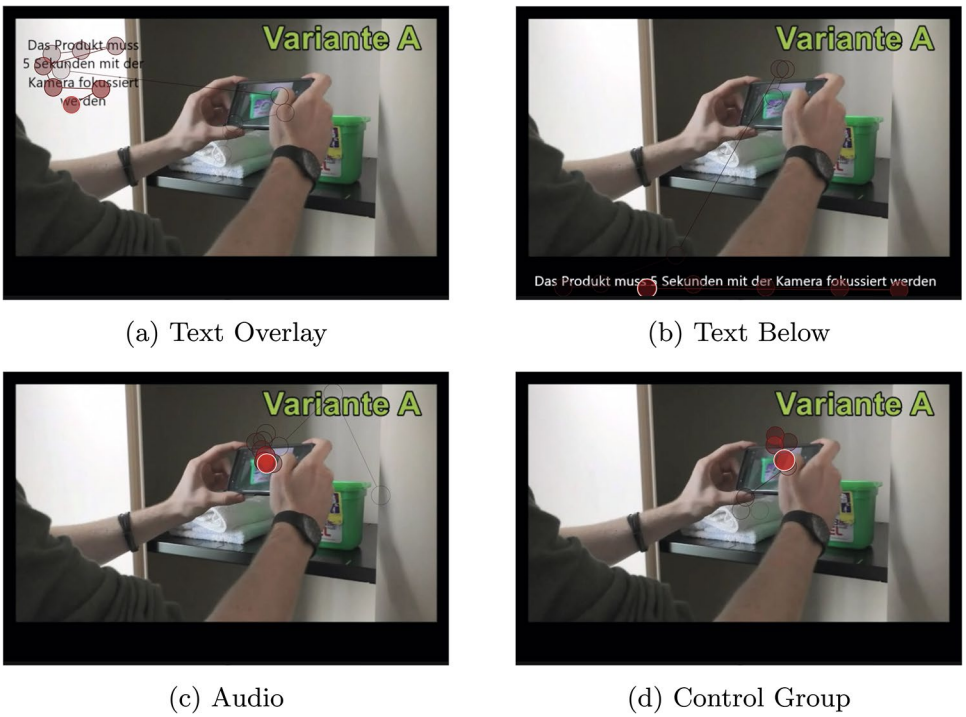
Fig. 2 An overview of the experiment design

content, demographics, questions specifically asking about participants' understanding of the additional information portrayed via text or audio and their opinions on the use of the respective method. For participants who were part of the control group the final parts of the questionnaire were not applicable as they did not receive any additional information. Instead, they were asked whether they felt that any information was missing. Figure 2 provides a simplified presentation of the groups participating in our experiment and its procedure.

3.6 Analysis procedures

Before interpreting our data, we carefully reviewed all recordings of the eye tracker to ensure that the gaze data could be properly captured at all times. This process led to the exclusion of four participants due to gaps in the data sets. These issues occurred most frequently for gaze points of participants reading the text below the video image. Additionally, we had to exclude a further participant as they did not complete the questionnaire. These participants are not described in the *Participants* section. We included only fully valid recordings in our results.

Figure 3 gives an example of one of the video scenes where additional information was added via text or audio. It shows how the added text below or on the video influenced the gaze behavior of the participants by representing fixations as red circles on the image. For the *TextBelow* variant,



(a) Text Overlay (b) Text Below

(c) Audio (d) Control Group

Fig. 3 Sample gaze plots of all groups for one of the information additions. The original video was produced by Schneider et al. [8]

Table 3 Results of eye tracking metrics for the whole videos

Metrics on video	TextBelow		TextOverlay		AudioVariant		Control	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Avg fixation duration (ms)	260.38	98.71	239.56	46.59	266.81	45.78	237.35	51.88
Fixation count	726.50	140.12	714.13	88.61	629.5	125.67	685.50	102.98
Saccade count	455.88	68.06	403.88	85.91	386.88	75.54	373.25	62.50

the gaze moved from the video content down to the textual information and for the *TextOverlay* group, the fixations shifted to the text on the image. On the other hand, for both the *AudioVariant* and the control group, subjects could keep focusing on the main content area of the video and did not get the visual distraction of the added text.

For questions about participants' understanding of the additional information, we strictly graded the provided answers in the categories *correct* and *incorrect*. Any answers that were partially correct or included inaccuracies were deemed *incorrect*.

Wherever tests for the statistical significance of differences were applicable, we tested our data sets for normal distribution using the Shapiro-Wilk test. Depending on the outcome of this test we chose between the t-test for normally distributed data and the Mann-Whitney U test when no normal distribution could be found.

Table 4 Results of eye tracking metrics for the additional text

Metrics on text	TextBelow		TextOverlay	
	Mean	SD	Mean	SD
Avg fixation duration (ms)	157	24.6	158	23.4
Fixation count	8.81	0.67	9.2	1.23
Text visit duration (s)	1.82	0.27	2.0	0.26

Table 5 Visit duration of overlay text area

Group	Overlay text area visit (s)	
	Mean	SD
TextBelow	0.003	0.007
TextOverlay	2.009	0.26
AudioVariant	0.044	0.099
Control group	0	0

4 Results

We present our results ordered by hypotheses.

4.1 Hypothesis 1

$H1_0$: There is no significant difference between the groups with additional information in viewing the videos in terms of visual effort.

Three different metrics were measured to gather information on the visual effort exerted by the participants. We recorded the average fixation duration, the fixation count and the saccade count for each of the variants. Table 3 provides an overview of these results for each group.

The data shows that participants watching one of the *TextVariants*, as well as the *AudioVariant*, showed a longer average fixation duration and had more fixations and saccades than those watching the original video as part of the control group. Higher numbers indicate a higher complexity and less efficient search for information [18]. I.e., our data indicates that the added texts and audio result in a slight increase in visual effort. However, these differences are not statistically significant.

When taking a closer look at the two *TextVariants*, we can examine four more eye tracking metrics for the text elements specifically. We compared the results of the two treatment groups watching the *TextVariants* in terms of their average fixation duration, fixation counts, visit durations, and visit counts of the text elements. Table 4 provides an overview of these text-specific metrics.

To validate our assumption that the text overlays were placed in video areas that usually did not include any valuable visual information, we compared the visit duration of the areas where the text was shown for all groups. The data can be seen in Table 5. It shows that all groups except for the *TextOverlay* group barely paid attention to these areas. This validates our assumption, but also shows that also the group seeing the text directly in the video image gets distracted from the actual relevant part of the video during those times.

After testing for normal distribution, we performed a t-test on the average fixation durations of the data obtained from participants watching the *TextVariants*. The test resulted in values of $t(14) = 0.09$ and $p = 0.93$. Therefore, we could not find a statistically significant difference between the two *TextVariants* regarding the average fixation duration of the text elements.

The same procedure was also performed for the fixation counts and text visit durations. The t-test resulted in values of $t(10.8) = 0.79$ and $p = 0.45$ for the fixation counts and

Table 6 Results of Shapiro–Wilk tests

Variant	W(8)	p	Normal Distribution?
TextBelow	0.777	0.023	No
TextOverlay	0.911	0.397	**Yes**
AudioVariant	0.870	0.161	**Yes**
Control Group	0.811	0.045	No

Boldened to indicate important results like a statistical significance or a confirmed normal distribution

$t(13.97) = 1.40$ and $p = 0.18$ for the text visit durations. Hence, our data does not present any statistically significant differences for these two metrics either. However, an initial tendency of higher reading times for the text overlays when compared to the texts shown below the video can be found.

In terms of the visit count, our data shows that all texts were looked at between one and three times. We found only a single occurrence of a text not being looked at by a participant. This might have been caused by more movement being present in the scene that was accompanied by this particular text.

All in all, we did not find any statistically significant differences between the treatment groups in terms of visual effort.

4.2 Hypothesis 2

$H2_0$: There is no difference between the four groups regarding the perceived cognitive load.

We asked participants to rate the level of mental effort they exerted while watching the vision video on a scale from 1 to 9. The resulting data was tested for normal distribution with the Shapiro-Wilk test before we performed analyses of the statistical significance of differences between the groups. The results of these tests for normal distribution are found in Table 6.

Depending on the presence or absence of a normal distribution we used different statistical tests. We used the t-test as it is robust against non-normally distributed data [41]. The Mann–Whitney U test was used as a nonparametrical alternative. We also applied the Bonferroni–Holm correction to reduce the threat of falsely accepting a positive result of a test for statistical significance. We used Bonferroni–Holm over the Bonferroni correction as a less conservative option based on the arguments laid out by Aickin and Gensler [48]. While the use of the Bonferroni correction is debated in literature [49, 50], we still chose to include the Bonferroni-Holm variation, as we deemed the threat of false positives to be too important for our findings. Table 7 gives an overview of these results.

4.3 Hypothesis 3

$H3_0$: There is no significant difference between the three groups with additional information in terms of the amount of newly gathered knowledge on the content of the video.

An overview of correctly or incorrectly answered questions regarding the content of the additional pieces of information portrayed via text or audio is presented in Fig. 4. Almost no difference can be seen between the three treatment groups. The group presented with the *TextBelow* variant answered 33 (68.75%) out of a total of 48 questions correctly. Both the group watching the *TextOverlay* variant with text overlays and the *AudioVariant* group gave correct answers to 32 (66.66%) questions. Therefore, we cannot reject $H3_0$. Our results display no difference between the three treatment groups regarding the amount of newly gathered knowledge on the content of the video.

Figure 5a presents an overview of the responses obtained from the treatment groups to a statement declaring that the *additional information was important*. On a 6-point Likert-scale ranging from 0 (do not agree at all) to 5 (fully agree), participants of the group looking at the *TextBelow* variant gave ratings of 4 six times with the remaining two responses giving a value of 5. For the group watching the *TextOverlay* variant, we obtained ratings of 2 and 3 once, three ratings of 4 and a further three ratings of 5. Lastly, the group watching the *AudioVariant* answered with the value 3 twice, with a single participant answering with a 4 and the five remaining participants indicating a value of 5. A similar general agreement was found regarding the statement that the *additional*

Table 7 Results for Hypotheses H2. Note that $H2_{TB,CG}$ would have been significant without the Bonferroni-Holm correction

Hypo.	Variant A	Variant B	p	p_corr	Reject $H2_{j,0}$?
$H2_{TB,TO}$	TextBelow	TextOverlay	0.142	0.710	No
$H2_{TB,AV}$	TextBelow	AudioVariant	0.430	0.760	No
$H2_{TB,CG}$	TextBelow	Control Group	**0.032**	0.192	No
$H2_{TO,AV}$	TextOverlay	AudioVariant	0.218	0.760	No
$H2_{TO,CG}$	TextOverlay	Control Group	0.342	0.760	No
$H2_{AV,CG}$	AudioVariant	Control Group	0.190	0.760	No

Boldened to indicate important results like a statistical significance or a confirmed normal distribution

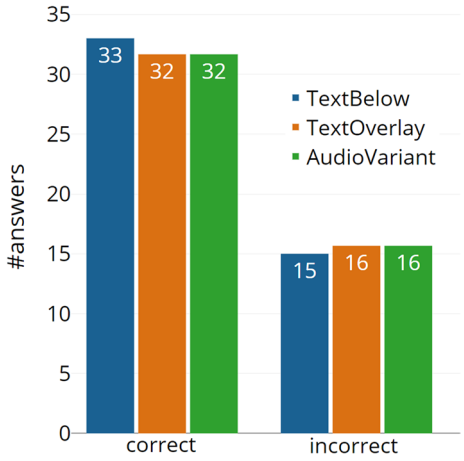


Fig. 4 Overview of correctly and incorrectly answered knowledge questions

information helped with comprehension. One single participant watching the *TextBelow* variant answered with a 3, while six other ratings of 4 and a final rating of 5 were also collected. The group receiving additional information via text overlays provided six ratings of 4 and a single rating of 5. Members of the group listening to the additional audio track responded with a single rating of 1, three ratings of 4 and three values of 5. An overview of these results can be found in Fig. 5b.

The questionnaire also included two statements designed to evaluate participants' opinions on the design of the text and audio elements. Both statements regarding the text elements were rated more favorably by participants of the treatment group watching *TextBelow*. The statement *I like the design of the texts* was rated with a single 0, 1, 2 and 3, with the remaining four responses indicating values of 4. The group watching the *TextOverlay* variant also gave rather diverse answers with two answers of 0, two answers of 1, three answers of 3 and a single answer with a value of 4. Similarly diverse responses were found regarding the statement *text position was appropriate*. Participants of the treatment group with text below the video gave two ratings of 2, three ratings of 4 and three ratings of 5. Answers from the group with text as overlays resulted in three ratings of 1, one rating of 2, 3 and 4 each and two ratings of 5. The group watching the *AudioVariant* answered similar questions in their questionnaire. The statement *I like the design of the audio information* was rated with two values of 1, two values of 3, a single value of 4 and three values of 5. Furthermore, the statement *the temporal positioning of the additional audio information* was answered with a single rating of 2, a single rating of 3, four ratings of 4 and two ratings of 5. An overview combining the responses of all treatment groups to the two statements can be found in Fig. 6a and b.

A further set of statements on which we collected our participants' opinions asked about the vision video itself. Figure 7 presents an overview of their responses. Once again, we employed 6-point Likert-scales ranging from 0 (do not agree at all) to 5 (fully agree). The first of these statements asked participants whether or not they liked the vision video. Our results show that the group watching *TextBelow* rated their liking of the video lower than the other two groups. All groups generally agreed with a statement regarding participants' opinion on whether the video provided important information. However, the group watching the *AudioVariant* indicated slightly lower values than their counterparts. Finally, we also collected participants' agreement with the statement *the video quality was sufficient* with which most participants agreed. As an additional question asked to participants of the control group, we inquired whether

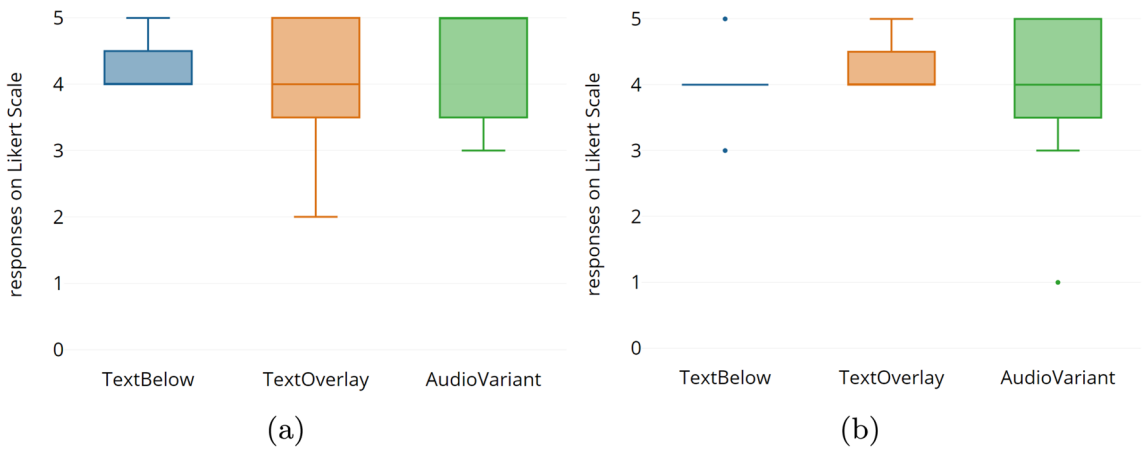


Fig. 5 Levels of agreement for the statements *the additional information was important* (**a**) and *the additional information helped with comprehension* (**b**) ranging from 0 (do not agree at all) to 5 (fully agree)

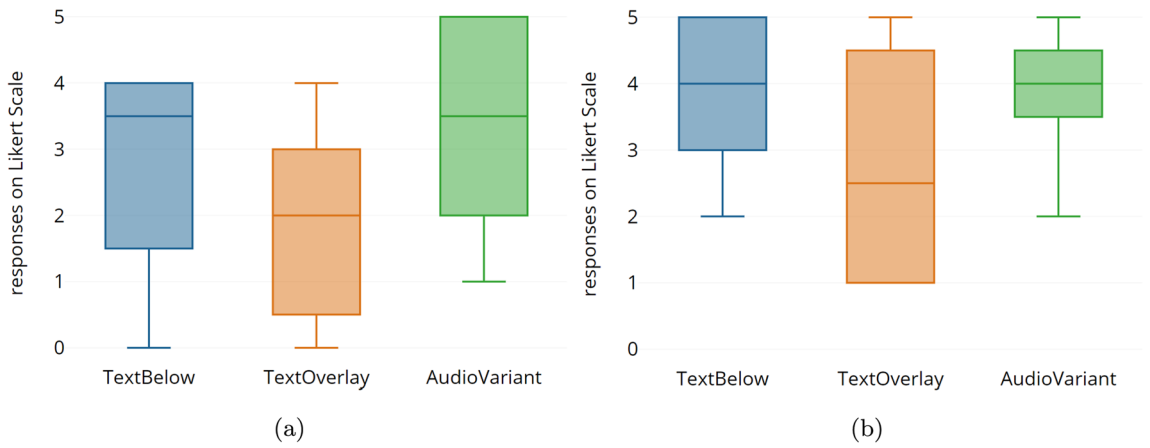


Fig. 6 Levels of agreement for the statements *I like the design of the texts* (**a**) and *the text position was appropriate* (**b**) ranging from 0 (do not agree at all) to 5 (fully agree)

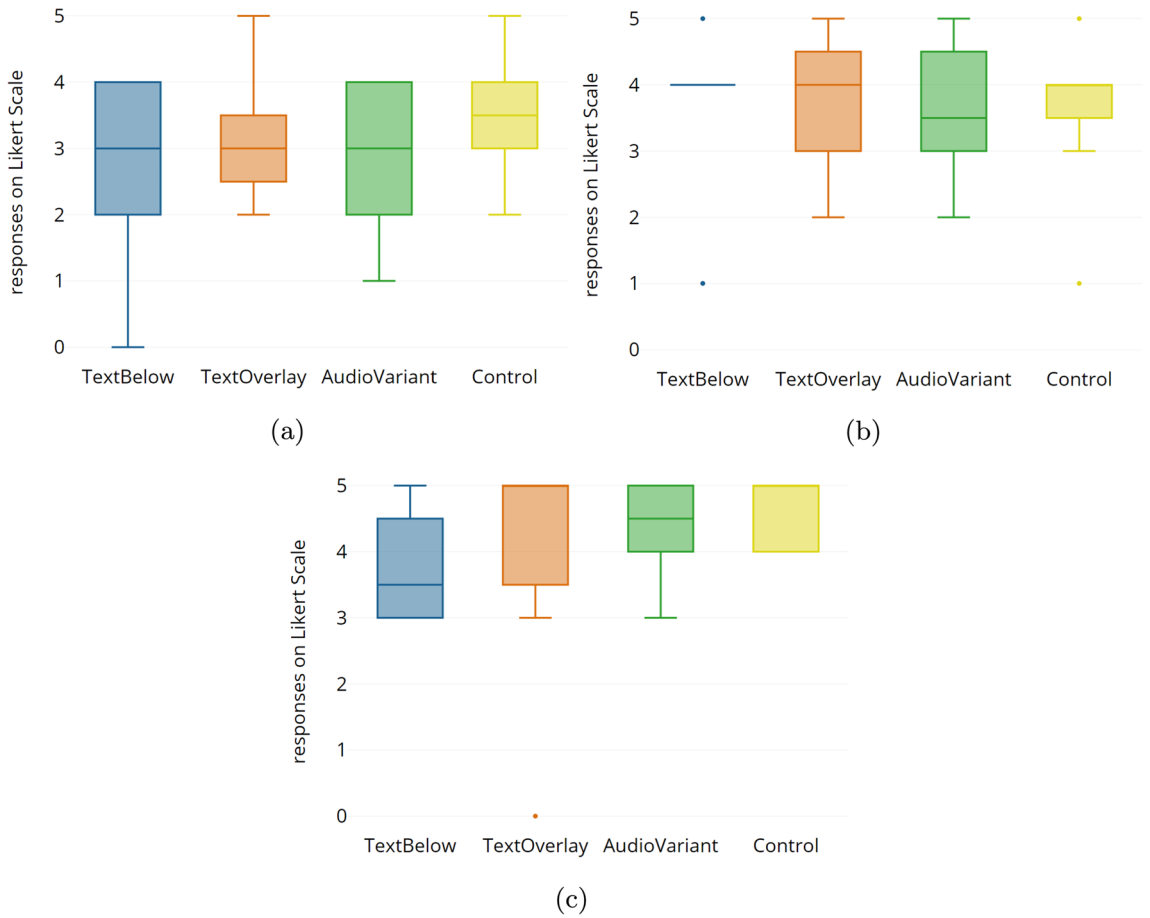


Fig. 7 Levels of agreement for the statements *I like the video* (**a**), *the video provides important information* (**b**) and *the video quality was sufficient* (**c**) ranging from 0 (do not agree at all) to 5 (fully agree)

they felt that they were missing any information that should have been included in the vision video. Three of the eight participants of this group indicated that they were missing information.

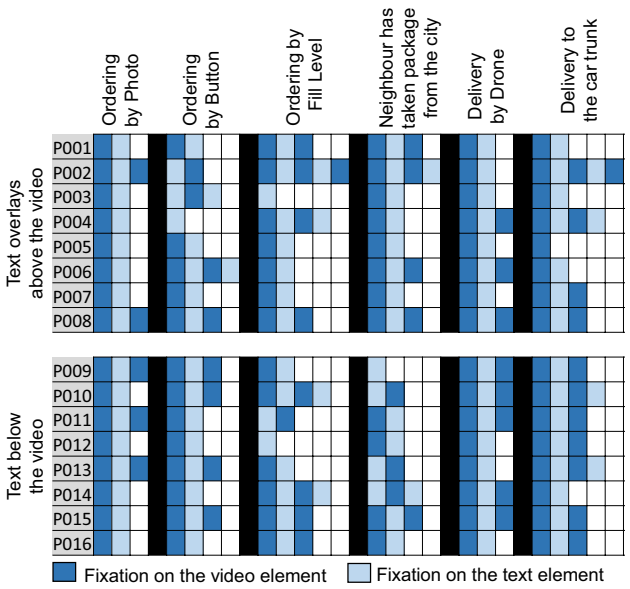


Fig. 8 Text-Video-Pattern

For the members of the treatment groups watching one of the *TextVariants*, we examined the eye tracking data in terms of the individual participants fixation patterns on two areas of interest. The first of these areas was the video content, while the second area was defined as the screen space containing the texts. We noted the fixations on these areas chronologically while watching the gaze recordings collected by the eye tracker. Figure 8 presents these patterns for both *TextVariant*-groups. Our data shows that each text was looked at at least once. One exception is present in a single gaze recording for the group viewing the variant with text overlays. Participant 5 did not fixate on the final text that was displayed.

4.4 Interpretation

Our study has lead to some interesting results. First of all, participants of the treatment groups were able to accurately reproduce 68.8% (*TextBelow*), 66.6% (*TextOverlay*) and 66.6% (*AudioVariant*) of the additional information. For the *TextVariants*, this finding is supported by the fact that all text elements were perceived and read by our participants. Only a single text addition out of a total of 96 (6 texts per video x 16 participants) was not fixated. Both *TextVariants* appear to have been perceived equally well. We found no significant difference between these two groups, neither in the reproduction of knowledge, nor in the fixation patterns we obtained in the eye tracking data. Other eye tracking metrics like the number of fixations and the fixation duration appear to be equal between the two *TextVariants* as well. We also found no significant difference in the reading time of the two text designs. However, a tendency of text overlays resulting in a higher reading time than the texts below the video could be observed.

When comparing the treatment group watching the *AudioVariant* to the other results, we obtained similar values for the eye tracking metrics. The recorded average fixation duration was the highest of all groups by a marginal, not statistically significant difference. In terms of the fixation and saccade counts the group watching the *AudioVariant* was the treatment group most closely comparable to the control group. The average fixation count calculated for the members of the group was the lowest result found in our experiment. However, the differences to the treatment groups once again were not statistically significant.

In terms of the cognitive load as self-reported by participants on a 9-point Likert scale [45], we found the highest ratings for the group with texts displayed below the video (MD = 4), followed by the *AudioVariant* (MD = 3), the group with texts as overlays (MD = 2.5) and finally without text (MD = 2). Participants indicated mostly low to medium values. After applying the Bonferroni-Holm correction, no statistically significant difference could be found between the groups in terms of their self-reported cognitive load.

Our questionnaire resulted in the finding that participants of all treatment groups evaluated the pieces of additional information as helpful for their understanding of the visions. Only a single participant disagreed with the corresponding statement. Our participants also indicated that the additional information was important for the content of the vision video. One single participant of the group with text overlays disagreed. Overall, there does not appear to be a significant difference between the three methods examined in our treatment groups. We summarize our findings by answering our research questions as follows:

Answer to research question 1:

The viewing behavior of viewers of vision videos is not majorly impacted by the inclusion of additional information via audio. Instead, it is influenced when the same information is displayed as text. We found no significant differences between the two *TextVariants*. Small differences between the treatment groups and the control group could be found in our data, however, these differences were not statistically significant either.

Answer to research question 2:

All three methods of adding information to vision videos add value to the understanding of the video content and enable viewers to expand their knowledge of the presented vision. All three methods are suited for the context of vision videos.

4.5 Threats to Validity

Although we had taken great care in the planning and execution of our study, its results are still threatened by some

aspects. We classify these threats according to Wohlin et al. [41].

4.5.1 Internal Validity

The internal validity of our results is threatened by the fact that participants theoretically could have simply guessed answers to content questions correctly. We implemented a strict grading of answers in the questionnaire by classifying inaccurate answers as false to minimize this threat. Additionally, the hygienic procedures performed due to the ongoing COVID-19 pandemic introduced a further threat to the internal validity. Facial masks that had to be worn by all participants could have impacted the eye tracking recordings. We ensured the absence of this factor by performing multiple test runs with different types of facial masks before conducting the main study.

Individual experiment sessions were supervised by a total of three different researchers which could have led to differences between the researchers impacting our results. We attempted to minimize this threat by carefully planning which information would be given out to participants and how questions would be answered. An additional threat stems from the use of two different experiment setups to allow two researchers to perform sessions simultaneously. The two setups were built to resemble one another as closely as possible. For example, we used the same screen sizes and eye tracker models for both setups.

We also recognize a threat to the internal validity of our results caused by the design of the blocks of texts in the *TextOverlay* variant. The color contrast between the text and the video content beneath it could have been suboptimal, which might have made it harder to read it. However, the low cognitive load reported by the participants viewing this video variant indicates that this threat did not meaningfully impact our results. We also looked to mimic a real world application of this method by keeping the effort required to implement the text as low as possible.

The internal validity of the results of participants watching the *AudioVariant* is also threatened by the audio quality of the video. Participants could have misheard the pieces of information provided by the narrator voice. However, no evidence of an impact of this threat could be found. When asked for feedback on the audio, some participants mentioned that they would have liked a slightly faster speech tempo, thereby indicating that there was no difficulty understanding the narrator. This statement also suggests that the audio clips were long enough to accurately portray the short pieces of additional information.

Lastly, the lack of statistically significant differences obtained from our study could have been caused by the sparseness with which the methods of adding information were used. However, we intentionally chose to add only a few pieces of information to the video presented in our study, as this is the intended use of our methods. The four ways of adding information to vision videos laid out in this paper should be used only when absolutely necessary.

4.5.2 External Validity

One threat to the external validity of our results is that the experiment included only short pieces of information. We also did not test different fonts or font sizes of texts and looked to keep the additional audio information similar in terms of tone and speech speed. Both the texts and the audio information were also visible or audible for just roughly three seconds. The generalizability of our results is diminished by these aspects. However, testing these parameters was not the intention of our study. We focused on the differences between the three variants tested in our study and the control group rather than looking for detailed information on these parameters.

Additionally, our groups consisted of only native speakers of the language of the texts. We therefore did not obtain any results regarding the potential difficulty of stakeholders working in their second language.

A third threat to the external validity are the demographics of our participants. Most of our participants were university students and of similar ages ranging from 19 to 29 years. Our participant selection was heavily reliant on an external platform that connects experimenters and university students who are looking to participate in scientific studies. An earlier search for participants without this external platform lead to an insufficient number of available test subjects. We accepted this threat to increase the sample size of our experiment.

4.5.3 Construct Validity

Our participant selection also threatens the construct validity of our results. No participant of our experiment was a real stakeholder. Therefore, they might not have cared about the contents of the presented vision. This could theoretically have lead to them being inattentive and less likely to give insightful answers. We found no evidence of this threat impacting our study results.

Furthermore, there could have been an impact on the data collection by the eye tracker due to personal characteristics of participants. For example, participants could have sat too closely to the eye tracker or might have been wearing glasses. To mitigate this threat, we calibrated the eye tracker for each participant and provided instructions on where exactly to sit in front of the screen according to the recommendations made by Tobii. We also observed participants on a second experimenter screen and reviewed all recorded

data to ensure data quality. We observed no impact on our eye tracking data from these threats.

Another aspect of our participant selection is their familiarity with subtitles. Participants who live in countries where movies are shown with subtitles might report a lesser impact of the *TextBelow* variant on their cognitive load. This could skew our data if only some participants are used to subtitles. We do not expect our results to be impacted by this threat as all participants came from the same geographic location. We also included the eye tracking metrics to obtain an objective measure of the visual effort.

The use of an external platform for the recruitment of our participants introduced a further threat to the construct validity. The platform required that participants were paid for their time. Our participants were therefore mostly motivated extrinsically. Once again, we accepted a potential impact of this threat in favor of a larger sample size and made clear that the financial incentive was in no way linked to the recorded performance but rather depended solely on the participation. Therefore, we do not expect a large impact of this threat on our results.

Moreover, the presence of experimenters during the sessions could have impacted our participants. Our careful planning of the experimenter's involvement in the study sessions included being as non-invasive as possible through being quiet and keeping responses to questions to a predetermined minimum.

4.5.4 Conclusion Validity

The conclusion validity of our results is threatened by the potential of technical issues leading to broken eye tracking recordings. We excluded any results obtained from participants with faulty eye tracking data. The results of such participants are not reported in this work.

Furthermore, the size of our sample also threatens its conclusion validity. We were unable to obtain a larger sample size due to the challenging circumstances of recruiting participants for an on-site experiment during the COVID-19 pandemic. Groups with more than eight participants each could have lead to more powerful results which might have presented statistically significant differences. This need for a larger sample size is further reinforced by the presence of a statistically significant result for $H2_{TB,CG}$ before the application of the Bonferroni-Holm correction. The use of the correction is debated in research [49, 50]. A larger sample size might lead to the discovery of a statistically significant result after the application of the correction.

However, we do not expect these statistically significant differences to result in unacceptable levels of cognitive load. This expectation is based on comments made by our participants over the course of the study which indicated that none of them were overwhelmed by the additional information. We are grateful to have acquired a total of 32 participants.

Another aspect limiting our sample size was that expensive eye tracking equipment and carefully chosen laboratory settings were required for all study sessions. The size of our study is in line with comparable experiments in other works using eye tracking [18]. We handled this threat by analyzing our data with statistical tests only where applicable and resorting to more qualitative analyses in all other cases.

5 Discussion

The results of our study have revealed some interesting insights on the potential of inserting additional information to vision videos via text or via audio. One of the most important findings of our experiment is that participants were able to accurately reproduce two thirds of the inserted additional information after watching the vision video only once. While one third of the answers to knowledge questions asked in the questionnaire were inaccurate or false, this result is still a positive indication for the value of the presented methods. The purpose of vision videos is not to enable stakeholders to reproduce every detail of the presented vision after watching the video for the first time. Instead, it is important that stakeholders are able to recognize misalignments between their own project vision and the one presented in the video. Participants generally seem to have recognized the pieces of additional information. The methods presented in this paper therefore are suitable ways of adding information to existing vision videos. Responses in the questionnaire regarding participants' opinions on whether or not the added information was important and beneficial to their comprehension of the project vision further reinforce this finding.

While we observed benefits of the methods of inserting additional information to existing vision videos, we also found results on their potential drawbacks. One example for such a drawback is a slightly increased visual effort exerted by our participants. The corresponding eye tracking metrics were increased slightly in all treatment groups when compared to the control group. A similar tendency can be observed for the self-reported cognitive load measures. However, we found no statistical significance in these differences. The fact that we found no statistically significant drawbacks between the control group and the treatment groups indicates that they are outweighed by the benefits of the presented methods. Even though it is possible, we do not expect a larger sample size to lead to significant differences. Nevertheless, a statistical significance might have been found with longer vision videos or a higher amount of additional information. However, we designed our evaluation with the recommended video length [9] in mind. The low to medium levels of cognitive

load indicated by participants of the treatment groups are reasonable for the duration of at most 5 min that was defined by the production guidelines for vision videos [9].

We recommend the use of the presented methods of inserting additional information to existing vision videos. Enriching vision videos with textual or auditory elements enables requirements engineers to include details that would otherwise take large amounts of effort to portray or might even be impossible to portray. We were unable to measure participants' understanding of the contents of the original vision due to the complexity of the presented variants. Participants were focusing on vastly different aspects of the vision which they thought to be most interesting. This meant that a binary definition of a correct or incorrect understanding of even parts of the vision was not applicable in our case. However, we found concrete information on the understanding of the additional pieces of information as these pieces were much shorter and more precise. The participants of our study were able to accurately reproduce two thirds of the additional information with only a negligible impact on their cognitive load. Nevertheless, we recommend keeping the additional pieces of information to a short length and to use them only when they are absolutely necessary. Information that can not be included as short texts or short audio clips should still be presented in written documents like a specification. Vision videos should be seen as supplementary material enabling a fast and direct communication of the project vision.

In addition to the general benefits of inserting additional information to vision videos, we also observed advantages and disadvantages of the individual methods.

The *TextBelow* variant is easy to implement for video producers and video editors as it can be added by simply creating a black bar beneath the video content and adding white text. This creates an aesthetic familiar to viewers due to the similarity with subtitles in movies. Adding text below the video most closely follows the *Affordable Video Approach* [8] and covers no parts of the video content. However, additional captions below the video might interfere with adding subtitles, e.g., for hearing-impaired viewers.

One of the main advantages of the *TextOverlay* variant is that the position of the text is closer to the video content. Viewers therefore have an easier time following the visual content in their peripheral vision while reading the texts. This can lessen the visual effort and the cognitive load compared to the *TextBelow* variant, as is evident in the results of our study. However, adding texts as overlays requires more effort from video producers and video editors. The overlays require a suitable position at which they do not obstruct any vital parts of the video and also demand a careful choice of the text color in order to maximize the readability on a dynamic background.

As for the *AudioVariant*, one of its main advantages is the fact that it adheres to the concept of the dual-channel assumption [34]. The visual aspects of the video can remain the same. This leads to results of the eye tracking methods close to those obtained for the control group. However, the additional audio track still increases the cognitive load experienced by participants. Furthermore, an implementation of the *AudioVariant* also requires considerable effort. While it is often easier to find a suitable temporal position for the audio clips than finding a suitable position for the text overlays of the *TextOverlay* variant, the *AudioVariant* still requires the recording of a narrator voice which introduces the need for suited audio equipment. Furthermore, audio tracks consist of a number of parameters like speech tempo, loudness and intensity that need to be kept in mind. Especially for longer audio clips, it might also get increasingly difficult to find suitable positions in the video without interfering with the original audio of the video.

Another option to recognize would be a combination of the *AudioVariant* with one of the *TextVariants*. All variants support the comprehension of the added information. A combination of audio and text could therefore result in the best comprehensibility. The drawbacks of this option are aligned with the ones outlined for the individual variants. Adding the pieces of information in two separate ways would further increase the effort required of video editors, especially since both a suitable moment in time for the audio track and sufficient screen space in the same video section need to be found. A temporal separation between an information being portrayed on an audio track and as text on the screen could lead to confusion. Another important aspect of the combination of multiple variants is the additional importance of the provided information perceived by viewers. This perceived importance could diminish the attention given to other aspects shown in the video at the same time. While this could be intended by the editors for especially important additions, we recommend vision video editors to be careful with the use of a combination of multiple variants.

All methods presented in this paper fulfill our goal of *finding a suitable method of inserting additional information to vision videos*. An analysis of the differences between the variants in terms of the visual effort and cognitive load revealed no statistical significance. We also did not observe a meaningful difference in participants' ability to reproduce the additional information. This indicates that all four methods manage to inform viewers of vision videos while introducing only negligible drawbacks to the viewing experience. The *TextBelow* variant appears to require a slightly larger increase of the cognitive load and visual effort from viewers than the other methods of inserting additional information. Nevertheless, its ease of implementation means that we recommend inserting text below the video as the most reasonable method.

The use of our methods can inform stakeholders on important aspects of the project vision. Therefore, all variants laid out in this paper can prevent requirements engineers from having to reshoot parts of a vision video if there is a need for clarification that arose only after the video had been created. We argue that especially the *TextVariants* should be seen as a meaningful tool to adhere to the *Affordable Video Approach* [8]. The use of all three of our methods can be relevant to the preproduction, when some details might be hard to visualize, and in the postproduction when unexpected needs for clarification are found. By supporting the vision comprehension, our methods can lessen the workload of requirements engineers while also improving the validation of visions and requirements in the requirements engineering process.

6 Conclusion

In this paper, we investigated three methods of inserting additional information into existing vision videos. These methods include the insertion of short texts below the video, short texts as overlays on top of the video content and the addition of short audio clips. All three methods were investigated in the context of an eye tracking study. The results of this study show that there are no statistically significant differences between the evaluated methods. Comparisons to a control group who watched the vision video without any additions also did not result in the discovery of any statistically significant differences. Participants were able to accurately reproduce most of the additional information. This finding is especially meaningful as an accurate reproduction goes beyond of what is required of stakeholders in the context of vision videos. Instead, stakeholders need to be able to recognize when the project vision presented in the vision video differs from their own. We argue that stakeholders are meaningfully supported in their task of unveiling misalignments between mental models by all three methods presented in this paper. This finding is further reinforced by our eye tracking recordings indicating that all texts were recognized. To adhere to the *Affordable Video Approach* [8], we recommend the insertion of additional information as texts below the video.

The methods presented in this paper support the comprehension of project visions presented in vision videos. They enable the presentation of feature details that would normally be hard to visualize while also replacing the need for potentially expensive reshoots when needs for clarification arise after the video production. Both the insertion of texts and the inclusion of additional audio clips should therefore be utilized as a suitable tool in the pre- and postproduction of vision videos. The use of our methods supports the achievement of the requirements engineering goal of creating a shared understanding between all relevant stakeholders.

In future research, we plan to consider deaf stakeholders. Textual information could enable them to discuss project visions using the same medium as other stakeholders. Additionally, future research could focus on how to collect the highest quantity or quality of feedback based on vision videos. For example, an interactive video player could be used to provide feedback to a project vision without having to switch to a different application. Another area of research is the application of vision videos in asynchronous communication contexts. Further research efforts could enable groups of stakeholders that are dispersed among different countries and time zones to discuss their project visions based on a vision video. This way, such research could enable the achievement of a shared understanding for more stakeholder groups and thereby aid in their software projects' success.

Acknowledgements This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under Grant No.: 289386339, project ViViUse. We would like to thank all our participants who took part in the study despite the corona pandemic.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

1. Creighton O, Ott M, Bruegge B (2006) Software cinema-video-based requirements engineering. In: 14th IEEE international requirements engineering conference (RE'06), pp. 109–118
2. Glinz M, Fricker SA (2015) On shared understanding in software engineering: an essay. Comput Sci-Res Develop 30:363–376
3. Norman DA (2002) The design of everyday things. Basic Books Inc, New York
4. Karras O, Schneider K, Fricker SA (2020) Representing software project vision by means of video: a quality model for vision videos. J Syst Softw 162:110479
5. Ambler S (2002) Agile modeling: effective practices for extreme programming and the unified process. Wiley, Hoboken
6. Karras O (2019) Communicating stakeholders' needs – vision videos to disclose, discuss, and align mental models for shared understanding. IEEE Software Blog. http://blog.ieeesoftware.org/2019/10/communicating-stakeholders-needs-with.html
7. Brill O, Schneider K, Knauss E (2010) Videos vs. use cases: can videos capture more requirements under time pressure? In:

International working conference on requirements engineering: foundation for software quality, pp. 30–44. Springer
8. Schneider K, Busch M, Karras O, Schrapel M, Rohs M (2019) Refining vision videos. In: International working conference on requirements engineering: foundation for software quality, pp. 135–150. Springer
9. Karras O, Schneider K (2021) An interdisciplinary guideline for the production of videos and vision videos by software professionals. Technical report, Software Engineering Group, Leibniz Universität Hannover. https://arxiv.org/abs/2001.06675v2
10. Schmedes M, Ahrens M, Nagel L, Schneider K (2022) Enriching vision videos with text: an eye tracking study. In: 30th IEEE International requirements engineering conference, RE 2022, Melbourne, Australia, August 15–19, 2022, pp. 77–87. https://doi.org/10.1109/RE54965.2022.00014
11. DeMarco T, Geertgens C (1990) Use of video for program documentation (experience report). In: Valette F, Freeman PA, Gaudel M (eds.) Proceedings of the 12th international conference on software engineering, Nice, France, March 26-30, pp. 126–128. IEEE Computer Society, Washington DC, USA (1990).http://dl.acm.org/citation.cfm?id=100314
12. Maiden N, Seyff N, Grunbacher P, Otojare OO, Mitteregger K (2007) Determining stakeholder needs in the workplace: how mobile technologies can help. IEEE Softw 24(2):46–52
13. Zachos K, Maiden N, Tosar A (2005) Rich-media scenarios for discovering requirements. IEEE Softw 22(5):89–97
14. Broll G, Hussmann H, Rukzio E, Wimmer R (2007) Using video clips to support requirements elicitation in focus groups-an experience report. In: SE 2007 workshop on multimedia requirements engineering
15. Busch M, Karras O, Schneider K, Ahrens M (2020) Vision meets visualization: are animated videos an alternative? In: Madhavji N, Pasquale L, Ferrari A, Gnesi S (eds) Requir Eng: Found Softw Qual. Springer, Cham, pp 277–292
16. Karras O, Polst S, Späth K (2020) Using vision videos in a virtual focus group: experiences and recommendations. Softwaretechnik-Trends 41
17. Nagel L, Shi J, Busch M (2021) Viewing vision videos online: opportunities for distributed stakeholders. In: 2021 IEEE 29th international requirements engineering conference workshops (REW), pp. 306–312
18. Sharafi Z, Soh Z, Guéhéneuc Y-G (2015) A systematic literature review on the usage of eye-tracking in software engineering. Infor Softw Technol 67:79–107
19. Ahrens M, Schneider K, Kiesling S (2016) How do we read specifications? experiences from an eye tracking study. In: International working conference on requirements engineering: foundation for software quality, pp. 301–317. Springer
20. Ahrens M, Schneider K (2020) Using eye tracking data to improve requirements specification use. In: Madhavji N, Pasquale L, Ferrari A, Gnesi S (eds) Requirements engineering: foundation for software quality. Springer, Cham, pp 36–51
21. Karras O, Risch A, Schneider K (2018) Interrelating use cases and associated requirements by links: An eye tracking study on the impact of different linking variants on the reading behavior. In: Proceedings of the 22nd International conference on evaluation and assessment in software engineering 2018, pp. 2–12
22. Karras O, Risch A, Klünder J (2021) Linking use cases and associated requirements: a replicated eye tracking study on the impact of linking variants on reading behavior. J Softw Eng Res Develop 9:5–15
23. Gralha C, Pereira R, Goulão M, Araujo J (2021) On the impact of using different templates on creating and understanding user stories. In: 2021 IEEE 29th International requirements engineering conference (RE), pp. 209–220. https://doi.org/10.1109/RE51729.2021.00026
24. Busjahn T, Schulte C, Busjahn A (2011) Analysis of code reading to gain more insight in program comprehension. Koli Calling '11, pp. 1–9. Association for Computing Machinery, New York
25. Jermann P, Nüssli M-A (2012) Effects of sharing text selections on gaze cross-recurrence and interaction quality in a pair programming task. CSCW'12. Association for Computing Machinery, New York
26. Goldstein RB, Woods RL, Peli E (2007) Where people look when watching movies: do all viewers look at the same place? Comput Biol Med 37(7):957–964
27. Srivastava N, Nawaz S, Newn J, Lodge J, Velloso E, M Erfani S, Gasevic D, Bailey J (2021) Are you with me? measurement of learners' video-watching attention with eye tracking. In: LAK21: 11th International learning analytics and knowledge conference, pp. 88–98
28. Brown A, Jones R, Crabb M, Sandford J, Brooks M, Armstrong M, Jay C (2015) Dynamic subtitles: The user experience. In: Proceedings of the ACM international conference on interactive experiences for TV and online video. TVX '15, pp. 103–112. Association for Computing Machinery, New York
29. Kruger J-L, Hefer E, Matthew G (2013) Measuring the impact of subtitles on cognitive load: Eye tracking and dynamic audiovisual texts. In: Proceedings of the 2013 conference on eye tracking South Africa. ETSA '13, pp. 62–66. Association for Computing Machinery, New York
30. Rayner K, McConkie GW (1976) What guides a reader's eye movements? Vision Res 16(8):829–837
31. Dogusoy B, Cicek F, Cagiltay K (2016) How serif and sans serif typefaces influence reading on screen: an eye tracking study. In: Marcus A (ed) Design, user experience, and usability: novel user experiences. Springer, Cham, pp 578–586
32. Rayner K (1977) Visual attention in reading: eye movements reflect cognitive processes. Memory Cognit 5(4):443–448
33. Hall RH, Hanna P (2004) The impact of web page text-background colour combinations on readability, retention, aesthetics and behavioural intention. Behav Inform Technol 23(3):183–195
34. Mayer RE (2021) In: Mayer, R.E., Fiorella, L. (eds.) Cognitive Theory of Multimedia Learning, 3rd edn. Cambridge handbooks in psychology, pp. 57–72. Cambridge University Press, Cambridge https://doi.org/10.1017/9781108894333.008
35. Mayer RE, Moreno R (2003) Nine ways to reduce cognitive load in multimedia learning. Educat Psychol 38(1):43–52
36. Schnotz W (2021) In: Mayer, R.E., Fiorella, L. (eds.) Integrated model of text and picture comprehension, 3rd edn. Cambridge handbooks in psychology, Cambridge University Press, Cambridge pp. 82–99. https://doi.org/10.1017/9781108894333.010
37. Simpson ML, Thomas KJ (1984) A comparison of oral and written text: a new perspective. Read Psychol 5(3):253–266
38. Software-Ergonomie : Empfehlungen für die Programmierung und Auswahl von Software, DIN-Taschenbuch. Beuth, Berlin (2004). https://www.tib.eu/de/suchen/id/TIBKAT%3A387737618
39. Richards JC (1983) Listening comprehension: approach, design, procedure. TESOL Quart 17(2):219–240
40. Basili VR, Caldiera G, Rombach DH (1994) The goal question metric approach, vol I. Wiley, Hoboken
41. Wohlin C, Runeson P, Höst M, Ohlsson MC, Regnell B, Wesslén A (2012) Experimentation in software engineering. Springer, Berlin & Heidelberg
42. Vegas S, Apa C, Juristo N (2015) Crossover designs in software engineering experiments: benefits and perils. IEEE Trans Softw Eng 42(2):120–135
43. Sharafi Z, Shaffer T, Sharif B, Guéhéneuc Y-G (2015) Eye-tracking metrics in software engineering. In: 2015 Asia-Pacific software engineering conference (APSEC), pp. 96–103. https://doi.org/10.1109/APSEC.2015.53

44. Jeanmart S, Gueheneuc Y-G, Sahraoui H, Habra N (2009) Impact of the visitor pattern on program comprehension and maintenance. In: 2009 3rd International symposium on empirical software engineering and measurement, pp. 69–78. https://doi.org/10.1109/ESEM.2009.5316015
45. Paas FG (1992) Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive-load approach. J Educat Psychol 84(4):429
46. Tuovinen JE, Paas F (2004) Exploring multidimensional approaches to the efficiency of instructional conditions. Instruct Sci 30:133–152
47. Nagel L, Schmedes M, Ahrens M, Schneider K (2023) Supplementary material - when details are difficult to portray: enriching vision videos. Zenodo. https://doi.org/10.5281/zenodo.7829952
48. Aickin M, Gensler H (1996) Adjusting for multiple testing when reporting research results: the bonferroni vs holm methods. Am J Public Health 86(5):726–728
49. Perneger TV (1998) What's wrong with bonferroni adjustments. Bmj 316(7139):1236–1238
50. Nakagawa S (2004) A farewell to bonferroni: the problems of low statistical power and publication bias. Behav Ecol 15(6):1044–1045