Eliciting Multimodal Approaches for Machine Learning-assisted Photobook Creation

Sara-Jane Bittner[1 [0009−0000−5171−7712]], Michael Barz[1,2 [0000−0001−6730−2466]], and Daniel Sonntag[1,2 [0000−0002−8857−8709]]

[1] German Research Center for Artificial Intelligence (DFKI)
{sara-jane.bittner,michael.barz, daniel.sonntag}@dfki.de
[2] University of Oldenburg

Abstract. Machine learning (ML) is increasingly applied in various end-user applications. To provide successful human-AI collaboration, co-creation for Interactive Machine Learning (IML) has become a growing topic, iteratively fusing the human creative view with the algorithmic strength to diverge ideas. Interactive photobook creation represents an ideal use case to investigate ML co-creation as it covers a range of typical ML tasks, like image retrieval, caption generation and layout generation. However, existing solutions do not exploit the benefits of introducing multimodal interaction to co-creation. We propose common operations for IML tasks related to interactive photobook creation and conduct an elicitation study (N=14) investigating which (combination of) modalities could well support these tasks. An open-ended questionnaire revealed how users imagine an ideal IML environment, focusing on device setup, key factors, and the utility of specific features. Our findings show that 1) enabling a wide variety of modalities allows for most intuitive interactions, 2) Informing users about uncommon modalities opens up suitable modality choices, that are otherwise missed, and 3) Multimodal interactions represent a high consensus, when chosen by the users.

Keywords: multimodal interaction · elicitation · machine learning · IML · user study · co-creation · interaction design.

1 Introduction

Machine learning (ML) is increasingly applied in a wide range of fields [12]. While ML models perform well in a range of applications, imperfect performance leads to users being unsatisfied and feeling reluctant to adapt the new technologies [17]. To match the model's performance the users' needs, it is crucial to enable a collaboration between the human view and the ML model. One promising approach is represented by interactive machine learning (IML), which has become a growing topic in the literature [30]: User input is used to retrain the model and improve the performance in following iterations [18]. In a further effort of fusing the human's point of view and algorithmic solutions, Artificial Intelligence (AI) co-creation was introduced into research. Human-in-the-loop feedback is

utilized in the collaborative human-AI process, in which the iterative creation of a human-centred artifact is set as an goal. Co-creation combines the strengths of humans - to act creatively and guide the process - with strengths of the ML - to explore possibilities and generate diverse artefacts [49, 35].

A key aspect of IML and co-creation is the interaction with the system. While most solutions for IML present mouse and keyboard interaction [54], Multimodal-Multisensor-Interfaces (MMI) enable intuitive and novel ways for the user to participate in the co-creation process. Research explores a wide range of modalities for user interactions such as speech, touch, gesture, and eye tracking [50, 40]. Including these interaction techniques can improve user experience and foster more dynamic, tailored feedback to the model, which then potentially improves model performance [24, 58, 32]. The selected modalities are especially important in IML co-creation as it influences the feedback characteristics, which directly impact the re-training of the model and the user experience.

One domain that represents a suitable testbed for co-creation and IML is photobook creation as it incorporates a set of complex ML tasks like image retrieval and selection, caption generation and designing of layout [47, 44]. While some studies already address individual tasks for ML-assisted photobook creation [33, 23, 2], they do not explore it in the context of a co-creation framework. Additionally, current solutions do not exploit the novel and intuitive possibilities of multimodal approaches. This reveals a gap in research for comprehensive multimodal interaction for IML-supported photobook co-creation.

Our paper investigates how humans use multimodal interaction intuitively in the context of ML-supported co-creation. For this, we apply the use case of interactive photobook creation as it covers a range of IML tasks. We hypothesize that further advances in IML will not only be based on technological factors, but that interactions with IML systems will benefit from the implementation of multimodal approaches in regards of model performance and user experience. After investigating the literature on IML co-creation and multimodal approaches, we conducted an elicitation study. In this study, participants have two rounds to go through a list of common operations (so-called referents) for interactive photobook creation and propose interactions. The proposed interactions are used to identify suitable modalities and combinations of such. In the first round, participants suggest modalities intuitively based on their experience. In the second round, they are primed with a common list of technologies in Human-Computer Interaction (HCI). Additionally, a questionnaire about ideal ML-assisted photobook creation is conducted, focusing on device setup and key factors. Based on this, our work covers the following contributions:

- Proposing a set of common operations in IML for tasks in photobook co-creation such as image-, caption- and layout-related tasks.
- Investigating users's perception of the commonly used modalities in HCI and explore what modalities users find suitable for which IML co-creation task.
- Investigating the users' thoughts about an ideal ML-assisted co-creation focusing on device setup and key factors.

2 Related Work

The following section covers ML and the use of human-in-the-loop feedback for interactive human-AI collaboration. Additionally, literature on multimodal approaches and the positive effect on intelligent user interfaces is presented. Then, photobook co-creation is introduced as an use case that combines different ML-supported tasks. Finally, elicitation studies are explained as a method to investigate intuitive interaction and human perception towards modalities.

2.1 Machine Learning

In recent years, the application of ML has increased in various fields such as material sciences, or data analysis [12]. Although ML models are widely used, they are black boxes that offer limited transparency to users. Furthermore, they often demonstrate imperfections and inconsistencies in performance, which can lead to user dissatisfaction and reluctance to adapt these technologies [17]. To improve on these imperfections, it is essential to combine the algorithmic performance with the human point of view. One promising approach is IML: Compared to traditional ML approaches, which do not allow for user intervention, IML enables model updates in response to user input [1]. User interaction with the model is often handled through the integration of human-in-the-loop feedback, which was shown to enhance the model performance [30, 43], adapt models to a specific domain [51], or tailor it to specific user preferences and needs [60]. By empowering users to influence the model's behaviour through experimentation, IML facilitates an intuitive and dynamic approach to model refinement [1, 16].

This human-in-the-loop feedback is often given through traditional mouse and keyboard interactions [54]. For example, a study increased the accessibility of text generation and segmentation proposals for videos by creating automatic generations that could be adjusted with mouse selection [60]. Based on the growing importance of IML, it is relevant to investigate effective interactions to interact with the user and collect feedback for the model effectively.

With recent advances focusing on generative AI, the field of IML has been extended to focus on the collaborative process of humans and AI. While traditional research in IML centres on optimising models through iterative user feedback, newer advances investigate co-creation as a process to generate new, user-tailored artefacts. Within this process, humans and AI are positioned as equal partners in a dynamic and interactive process: Humans contribute intuition, guidance, and flexibility, while AI assists by exploring possibilities, generating diverse artefacts, and supporting iterative development; this collaboration empowers users to actively steer the creative process, leveraging AI to achieve outcomes that surpass what either could accomplish alone [42, 57, 35, 49].

2.2 Multimodal-multisensor Interfaces

IML can be combined with MMI to enable intuitive and novel interaction approaches. MMI utilize various human senses and behavioral cues—such as speech,

mid-air gestures, touch, controller and gaze—to enhance the interaction between user and system [50, 40]. In the context of developing multimodal interaction design for ML-supported systems, the concept of Intelligent User Interfaces (IUI) was introduced [34]. They define IUI as "human-machine interfaces that aim to improve the efficiency, effectiveness, and naturalness of human-machine interaction by representing, reasoning, and acting on models of the user, domain, task, discourse, and media." A combination of HCI and advances in technology in a synergistic manner will lead to benefits for users. With that, it aligns with Jameson et al.'s metaphor of a *binocular view* [27]: It introduces the design of interactions and intelligent algorithms as a single design problem, in which both aspects should be considered simultaneously for successful development.

By incorporating multiple input modalities, MMI can improve flexibility [58], boost user satisfaction, and enhance task efficiency during model interaction [22], compared to traditional mouse-and-keyboard interactions. For example, multimodal interactions - like eye tracking gestures, and natural language - increase time efficiency and are effective for user satisfaction [22]. The impact of the selected modalities on the interaction is especially important in IML co-creation as it can influence the feedback characteristics, which directly impacts the retraining of the model and the user experience.

Researchers have investigated the effect of a range of MMI, including speech, eye tracking, controller, touch and mid-air gestures: *Speech-based interfaces* show two main benefits: Users experience a lower barrier to express their intent [3], and enable flexible, natural language interactions [13]. For example, integrating a multimodal approach in which users could specify and perform semantic image search tasks can enhance image retrieval [4]. The user provides natural language queries as well as positive and negative examples, and based on them, fitting images are retrieved. *Eye tracking* enables the user to engage with elements based on their gaze with hands-free, intuitive interactions [25, 15]. Through that, assumptions about the intention and ongoing actions can be made. This can be important to indicate relevant context information for spoken feedback as humans fixate on an object just before they include a speech command [21]. The use of *gestures* represents an intuitive way to interact with a system [59]. It was applied in a range of applications as an interaction modality and can improve the user experience [32, 36, 24]. *Controllers* pose a similar function to gestures and can be used for navigation, interaction, and to provide context. Integrating multimodal, natural interaction methods into IML and co-creation processes has the potential to enhance the engagement, adaptability and flexibility of human-AI partnerships. However, IML remains largely confined to 2D interfaces and conventional input methods such as mouse and keyboard [54]. We are not aware of studies that investigate which modalities or combinations of such could be effective and are intuitive for human-in-the-loop co-creation with ML systems.

2.3 UseCase: Photobook Creation

The domain of photobook creation presents an opportunity to apply co-creation and IML due to its incorporation of complex ML tasks like image retrieval and

selection, caption generation and designing of layout [44, 47, 48]. Although individual ML tasks have been investigated—such as image selection techniques [33], user-specific caption generation with iterative feedback [2], and multimodal interactions for person identification [23]—there's a lack of research within a co-creation framework. Existing solutions like PICANOVA[3], photobook.ai[4], and Journi[5] focus on automating the process towards a final product, but they lack the iterative and flexible nature of co-creation and are constrained by their commercial ties [19, 29, 8]. Additionally, most available solutions apply traditional mouse-and-keyboard interactions, not utilizing the indicated benefits of multimodal interaction for co-creation. This reveals a gap in research for comprehensive multimodal human-AI collaboration for IML-supported photobook co-creation. MMI can be utilized in various ways for co-creation tasks. For instance, combining speech and eye tracking could be used to indicate the objective of user reference for a caption-feedback: the user has an image in which their dog *Paula* is sitting in front of a lake. The system displays the caption *A dog sitting in front of a lake*, missing the context information of the dog's name. Then the user could fixate the dog and explain *This is my dog Paula*, which can then be processed and corrected by the system. Adapting to the specific user's context is especially relevant in photobook creation as important information can vary between users and the motivation of the photobook significantly [5].

2.4 Elicitation Studies

Elicitation studies have gained popularity for the design of natural interfaces and new symbol sets for the use of such [56]. They represent a method of participatory design in which the user group is actively involved in the design process to create a system that fits the user needs better [46, 7]. In regards to elicitation, the immediate use of a system can be improved by increasing the guessability of the input that the system requires to execute a certain operation [55, 24]. Early studies mainly focused on gesture elicitation [55, 56]. However, newer advances tend to investigate multimodal interaction [36, 24]. For example, Morris [36] investigated the use of gestures and speech for interacting with a web browser on a TV. The study highlighted how users would be open to using their TV for web browsing, and that multimodal synonyms should be implemented so that the user can decide which modality they would like to use for the interaction. Further, a study by Herbig et al. investigated the interaction of post-editing machine translation systems with 5 modalities ranging from mouse and keyboard to eye tracking and a combination of these [24].

To execute an elicitation, the participant gets a list of referents, which are common operations in the given system. For example, a referent for editing a caption could entail: *"Caption Location: The caption is to general. You want the caption to contain the specific place."* For each referent, the participant proposes one or several actions containing how they would achieve these referents [20].

[3] www.picanova.com
[4] https://photobook.ai/
[5] https://www.journiapp.com

An elicitation study often contains two rounds: 2) An unbiased elicitation - in which the referents are iterated through without any additional information. 2) A biased follow-up, in which participants receive a list of modalities. That way, awareness is created for technologies that might be uncommon in the daily life of participants, and which they hence might overlook in the task. For each modality, an explanation is provided, and an example is given of how the modality could be used in an application. For example, speech is connected to voice assistance, and touch is connected to smartphone usage.

The participants are introduced to two rules before proposing actions: 1) The participants' suggestions are always acceptable - This rule enables intuitive and unlimited proposals [56] and 2) The system will recognize the participants' proposals correctly without any technical errors - This rule leads participants away from technical thinking to not let them be limited by current technical capabilities [38]. Further, potential biases need to be considered: The legacy bias describes that the user's previous experience with technology has an impact on their proposed actions [36]: Technologies that participants are more familiar with are proposed in higher number and are perceived more positively. In comparison, technologies that participants are less familiar with, could be missed, as participants potentially do not recall them as an option during proposals. Approaches to limit this bias include proposing in groups (*partners*), or proposing several actions (*production*) for one referent. Additionally, as applied in the biased elicitation *priming* the participant by introducing a set of technologies before the elicitation, can open up the participant's mind to unfamiliar technologies [36].

Different formalization strategies are applied for elicitation studies which differ in the amount of proposals per referent: For elicitations that only allow for one proposal per referent, the *agreement rate* - which indicates the consensus between participants - and *co-agreement rate* - which indicates how much agreement two referents share- are applied [36]. In our study, participants are encouraged to propose multiple interactions, aligning with the production strategy [36, 24]. Therefore, we apply the *max-consensus*, which represents the percentage of participants proposing the most popular proposal, and the *consensus-distinct-ratio*, which represents the percentage of distinct interactions for a given referent, when this interaction reached a threshold.

3 User study on interactive Photobook Co-Creation

We conducted a user study to investigate the users' perception on different modalities in co-creation and to explore which multimodal approaches might be effective for photobook editing in a human-AI collaboration. We aim to derive implications for multimodal interaction with ML systems, including ML tasks such as image selection, caption generation and layout design. An ethical approval was submitted and approved at the *DFKI Ethics Committee*. Participants were recruited through distributed flyers at photobook creation booths, or online through a mailing list. Participants were a minimum of 18 years old and had experience creating at least one photobook with a web or desktop service.

3.1 Study Plan

The following section introduces the study plan. For this, the participants, as well as the procedure with corresponding methods and measures, are introduced.

Participants In total, 14 user participated in our user study, who were between 22 and 67 years old (M=33, Std=9.60). Nine participants are female (64%) and five participants are male (36%). They have created a minimum of one and a maximum of ten photobooks (M=2.67, $Std.$=2.56). The participants reported how familiar they are with 1) devices and 2) modalities in five categories: *Never*, *less than once per month*, *between once a week and once a month*, *a few times per week*, *almost every day*. First, the modalities included gestures, speech, eye tracking, controller, and touch. Participants are most familiar with touch, which was used by 13 out of 14 participants daily. The participants are least familiar with gestures and eye tracking. For each 10 of the participants reported to have *never used it*. Speech and controller were used in a medium range with some participants *never* or *rarely* using them and some participants using them *between once a week and once a month*. Second, the devices included computers, virtual reality (VR), situated screens and mobile handhelds such as tablets or smartphones. The participants were most familiar with the computer and handheld devices, with 13 of 14 participants that use them daily. However, participants were mainly unfamiliar with VR environments and situated screens: Seven out of 14 participants have *never* used VR and 13 out of 14 participants have *never* used a situated screen. Lastly, their self-assessed experience in creating photobooks averaged M=3.5 ($Std.$=0.50) on a scale from 1 (unexperienced) to 5 (very experienced), indicating they assume that they are quite experienced. In general, they assessed the creation of a photobook to be a rather easy task (M=2.28, $Std.$=0.88, 1=very easy, 5=very hard).

Procedure First, participants filled out a consent form and received a briefing with the study context. Then, the study plan consisted of three steps: 1) An initial questionnaire consisting of questions about their prior photobook experience. 2) The elicitation task, which can be split into an unbiased and a biased elicitation. 3) A final questionnaire, that focuses on ideal photobook creation.

1) Initial Questionnaire First, a questionnaire that covers the user's experience with photobook editing is filled out. It features pain and gain points, and motivation for photobook creation. Additionally, demographic data is gathered.

2) Elicitation Similar to [24], we conducted a study with an unbiased and biased elicitation. Aligning with previous studies, we provided a non-interactive prototype, which is held simple to not limit the interaction space of the participants [24]. We used FIGMA to provide a low-fidelity version, which displayed a general photobook view (see figure 1). In the beginning participants were informed to extend the shown prototype by their own expectations of the system. Participants were able to express these during the proposals.

First, the participants are introduced to the concept of multimodal photobook creation. Corresponding to the common guidelines for elicitation studies, the participants were briefed that "their proposed interactions will be recognized and executed as they intended perfectly" and that they should "not propose mouse and keyboard interactions". While, traditional interaction with mouse and keyboard, might be efficient, our paper aims to explore AI co-creation in a novel interaction space, considering different modalities and combinations aligning with previous work [36, 24]. After this, participants will go through a list of 14 referents, and will propose several actions for each. The referents represent common operations that are relevant for the process of photobook creation [20].

To derive referents, the process of photobook creation was analysed using four photobook services. Based on this, 14 common operations were defined. For each referent, a description was presented to the participant to ensure understanding, which can be seen in table 1. The referents focus on three common ML tasks: Selection of images, creation of captions, and adjustment of the layout. The referents were adapted to the use case of ML co-creation, with the majority of referents posing the potential for AI collaboration. Referents one to five cover image-related tasks. For example, referent *4) image search* covers the retrieval of images based on the processing and with that gained understanding of the image space. Further, referents six to eleven cover caption-related tasks, and, referents 12-13 refer to layout-related tasks. In the context of ML co-creation, a majority of the referents focus on users changing the current photobook version via feedback to the system. In the study they are ordered based on a balanced Latin square. The participants are encouraged to propose more than on action per referent. This addresses the legacy bias, in which participants tend to propose interactions that they are familiar with [37, 6].

After the unbiased elicitation, the second round represents a biased elicitation: Participants are introduced to a list of six modalities and their descriptions: Mouse and keyboard (MK), touch (T), mid-air gestures (G), speech (S), eye tracking (E) and controllers (C). Also, combinations of modalities were listed as combinations of the letters, such as *SE* for speech and eye tracking (See table 4). With that, we introduce a priming strategy into the biased elicitation. This opens up the proposals to modalities that participants have not considered intuitively during the prior round of proposals [24]. To not limit the possible interaction space and enable participants to propose interactions intuitively, we decided to include a wide range of modalities, including modalities that are considered for more passive interaction such as eye-tracking [40]. For each modality, a description and an example video was shown, depicting how the modality can be used. A correct understanding of the modalities is crucial for an elicitation study. Previous approaches utilized mainly textual descriptions, in which the functionality of modalities might be hard to grasp [24]. Videos show the use in an interaction context, leading to a better understanding. If possible, the video corresponded to the use of the modality in daily life. For example, speech was connected to voice assistance and touch to smartphone usage. To make sure that participants had sufficient understanding of each modality, they were asked to propose an

interaction per modality for an example referent. After this, participants went through the list of referents a second time, updating their proposals.

Table 1. Referents (*Ref*) of the elicitation study.

Nr	Referent	Description
1	insert image	Add a picture at a specific location.
2	image selection	The system shows you an overview of images and you want to select some for the photobook
3	image alternatives	You want to get image alternatives to replace the current image.
4	image search	You want to replace the current image with another image. You know which picture you are looking for.
5	similar images	There is a better version of the current image (better crop, image definition). You want to replace the current image with it.
6	caption alternative	You want to display alternative captions and select one
7	caption style	The caption is too objective. You want a funnier caption.
8	caption context	The caption is too general. You want it to contain more contextual information.
9	caption location	The caption is too general. You want the caption to contain the specific place.
10	caption person	The caption is too general. You want it to contain a specific person that is in the image.
11	caption correction	The statement in the caption is incorrect. You would like to correct it. (*Example: "A farm is surrounded by water" rather than "A castle is surrounded by water"*)
12	move elements	You want to move elements to another position on the page.
13	change image size	You want to change the size of an image.
14	change font size	You want to change the size of the caption.

3) Ideal Multimodal Photobook Creation Finally, a second questionnaire is conducted to investigate how participants would picture multimodal photobook creation ideally. Hereby, they were asked how they imagined interactive photobook creation in different settings such as virtual reality (VR), situated screens, mobile applications and web or desktop applications. Further, they were asked how such a design would look, how they would rate the utility of features and components, and how they imagined interacting with them. As an analysis method for the questionnaire, a *reflexive thematic analysis* (RTA) was conducted. Braun and Clarke's analysis was first published in 2006 and has been widely validated to be a suitable method of qualitatively analysing data sets [9, 10]. It was chosen based on the importance of reflexivity in qualitative studies. An *RTA* consists of 6 steps that are executed incrementally (See appendix 7). Lastly, participants were asked to rate how often they use the proposed modalities.

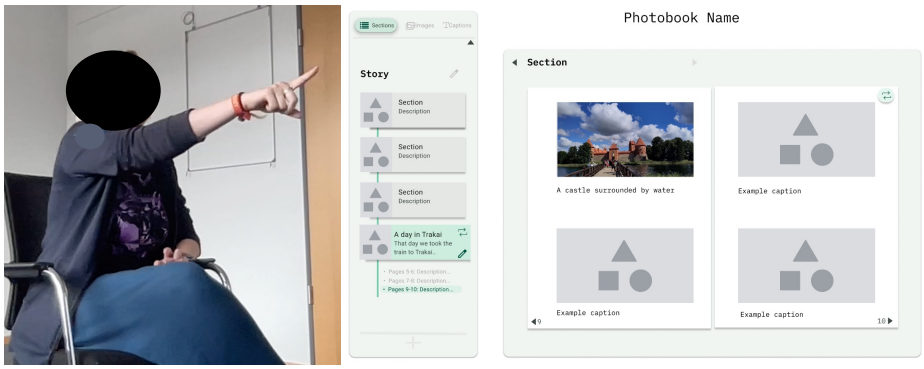

Fig. 1. Overview of the elicitation study displayed with re-enacted participant left and screenshot of the photobook mock up right.

3.2 Results

The results of our user study are presented below. 1) A overview of the proposed modalities that were utilized by participants is presented. 2) More detailed insights per proposed modality are given, focusing on the use of certain ML-supported tasks and in comparison of the unbiased and biased elicitation. 3) Insights into different setups for ML co-creation and highlighted key factors that were derived from the questionnaire are reported. Concerning data saturation, 14 participants suffices the amount to draw conclusions: 1) Regarding the elicitation task, a literature review reports 38% of elicitation studies to include in between 10 to 20 participants [53]. Thus, 14 participants, align with previous approaches and allow to draw conclusions about the proposed interactions [24]. This aligns with the assessment of the semi-structured interview as in the analysis new themes stopped emerging in responses.

Elicitation In the unbiased elicitation, participants suggested a total of 172 common proposals. After the priming strategy was applied in the biased elicitation, that number increased by 39.5% to 240 common proposals. For both elicitation tasks, the most proposed modality was speech (unbiased:86, biased:91), followed by touch (unbiased:56, biased:71) and a combination of speech and touch (unbiased:20, biased:40). Also, during the biased elicitation, participants proposed gesture (17), speech combined with gesture (15) and eye tracking (6). For the unbiased elicitation, speech and touch seem to be similarly popular for image-related tasks, while speech seems to be preferred for caption-related tasks, and touch for layout tasks. After applying the priming strategy in the biased elicitation, not only speech was considered for caption-related tasks but also the combination of speech with 1) touch or 2) mid-air gestures. In general, after priming mid-air gestures show the highest increase of proposed interactions.

In the unbiased elicitation, 13.7% of proposals are multimodal, with the biased elicitation showing a small increase to 14.6%. While this is considerably

higher than for Morris's [37] study (3.1%), it only reaches less than half of the amount of multimodal proposal compared to Herbig et al.'s [24] study (33.0%), who investigated multimodal approaches for post-editing in machine translation.

For each referent, the max-consensus ratio - the peak - and the consensus-distinct ratio - the distribution of proposals - were measured. Two different granularities for these measures were calculated: The ratios 1) overall modalities (All C_m, C_d), and 2) individually for each modality (modality C_m, C_d). In later, the ratios give indications of the different interactions proposed for one modality. For example, verbally commanding to add images via number (*"Add image 1."*) versus verbally commanding to add images via description (*"Add images with a sunset"*). For this study, only modalities that got ≥ 3 proposals were included. The results can be seen for the unbiased elicitation in tables 2 and 5 and for the biased elicitation in tables 3 and 6. Table 5 and 6 display extended results and can be found in the appendix. Regarding different modalities, *speech* shows the highest *max-consensus* and *consensus-distinct ratio* for caption-related and image-related tasks. For both unbiased and biased elicitation, referents with the overall highest ratio were the caption-related referents *caption alternatives*, changing *caption style* and *caption correction*, followed by the image-related referent *similar images*. However, after priming, speech was additionally proposed in combination with touch or mid-air gestures. For layout-related tasks, *touch* showed the highest measures for both rounds of elicitation.

Table 2. Unbiased Elicitation: Proposals per referent (Ref). Number of proposals (total and distinct), the percentage of multimodal proposals (MM%), and modalities suggested ≥ 3 times.

Nr	Referent	Number (tot/dist)	MM% (tot/dist)	Common Proposals
1	insert image	18/5	5.5/20	T(8),S(4),ST(3)
2	image selection	17/5	5.8/20	T(9),S(4)
3	image alternatives	14/4	0/0	S(8),T(4)
4	image search	18/8	22.2/50	S(10)
5	similar images	16/5	6.25/20	S(9)
6	captions alternative	17/7	17.6/42.8	S(6),ST(4)
7	caption style	14/4	14.2/50	T(8),S(3)
8	caption context	17/5	17.6/60	S(11),ST(3)
9	caption location	16/5	18.8/60	S(9),ST(3)
10	caption person	17/5	23.5/80	S(9),ST(4)
11	caption correction	15/6	26.7/66.7	S(8),ST(3)
12	move elements	16/5	6.2/20	T(9),S(3)
13	change image size	15/6	13.3/33.3	T(9)
14	change font size	14/5	14.3/40	T(9)

Speech Participants showed a high experience with speech compared to other modalities, with 14.3% using it daily, 21.4% multiple times per week, and 28.6%

Table 3. Biased Elicitation: Proposals per referent (Ref). Number of proposals (total and distinct), the percentage of multimodal proposals (MM%), modalities suggested ≥ 3 times.

Nr	Referent	Number (tot/dist)	MM% (tot/dist)	Common Proposals
1	insert image	26/6	7.7/33.3	T(11),S(4),ST(4),E(3)
2	image selection	21/5	47.6/20	T(11),S(4),G(3)
3	image alternatives	19/6	5.3/16.7	S(9),T(5)
4	image search	21/7	14.2/42.9	S(11),ST(4)
5	similar images	20/5	5/20	S(9),T(5),ST(3)
6	captions alternative	24/6	12.5/50	S(6),ST(6),SG(4),T(4),E(3)
7	caption style	19/4	10.5/50	S(8),T(4),ST(5)
8	caption context	19/5	15.7/60	S(11),ST(4)
9	caption location	19/5	15.7/60	S(9),ST(4),SG(4)
10	caption person	21/5	19.0/80	S(9),ST(5),SG(4)
11	caption correction	21/6	19.0/67	S(11),ST(4),SG(3)
12	move elements	25/6	8/33	T(11),G(6),S(3)
13	change image size	22/7	13.6/42.9	T(10),G(6)
14	change font size	19/5	10.5/40	T(10),G(5)

at least once a month. In both rounds of elicitation, Speech was proposed for all referents other than the layout-related referents *13) change image size* and *14) change font size*. Speech was proposed in combination with a command and a specification. Two different strategies for co-creation can be distinguished: 1) Relying on the system's "intelligence" - the changes are left to the system choice. This includes requesting a new image with the specification of processed image data (P1: *"Show me similar images"* or prompting the system to consider metadata (P6: *"Please include the location in the caption"*). 2) Specifying the exact changes that the system should make. This includes specifying a description of an image (P4: *"Find me the picture of the beach with an ice cream"*) or commanding an edit of the new caption (*P12"Insert: Maria in front of Trakai Castle"*)

For both rounds, the *max-consensus ratio* and *consensus-distinct ratio* are rather high. Regarding caption- and image-related tasks, the *consensus-consensus ratio* for the unbiased rounds showed slightly higher values with 75% for all but one referent. Compared to that, the biased elicitation indicates a little lower consensus over all referents featuring only speech (63.7% to 100%), which indicates that most people would use speech in a similar way. The *consensus-distinct ratio* for image- and caption-related tasks range from 0.33 to 1.00, indicating that while a few people deviate, the majority often choose similar speech interactions.

Touch Participants were most familiar with using touch, with 100% reporting to use it daily. It represents the modality with the second-highest number of proposals in the unbiased and biased elicitation. It was mostly proposed for image- and layout-related tasks. For caption-related tasks, touch was only considered for referent *7) caption style*. The pointing touch-gesture was proposed most: Shortly pointing on an element indicates selection, while longer pointing initiated further

interaction with a sub-menu. Also, a pointing touch-gesture was commonly used for drag and drop. Specifically for changing the size of elements for the layout tasks, two strategies were used: Pinching the index finger and thumb together and diagonally moving them apart or closing together, or selecting the corner of an element and dragging it to another size. Specifically, for font size, some people used touch to change the numerical value of the font size, as it is common in word editors. Most unity is shown for the referent *5) similar images* where all participants proposed to touch longer on an image to receive similar suggestions.

For both rounds, the *max-consensus ratio* is medium to high. Regarding image- and layout-related tasks, the unbiased rounds showed slightly lower values(50% to 87.5%) compared to the biased round (50% to 90.1%). This indicated a medium consensus of proposed touch interactions between the participants. The consensus-distinct ratio for the image-related tasks ranges widely for the unbiased (0.00 to 0.66) and biased elicitation(0.00 to 1.00), which indicates that participants used a wide variety of touch interactions. In comparison, the consensus-distinct ratio for the layout-related tasks ranges from 0.50 to 0.75 for both rounds of elicitation, indicating that while a few people deviate in proposed touch interactions, there is a fair overlap of similar intuitive interactions.

Speech and Touch Participants were quite familiar with speech and touch as individual modalities (see the results of speech 3.2 and touch 3.2). For the unbiased elicitation, the combination of speech and touch (ST) was mainly used for caption-related (excluding referent *7) caption style*) and once for image-related tasks with referent *1 insert image*. However, in the biased elicitation the number of proposals for image-related tasks including ST increased. This combination was mostly proposed by participants who have suggested speech-only interactions in the unbiased elicitation. For both elicitations, the participants used mainly two approaches for ST interactions: 1) Pointing at the element and then specifying the aimed action with a command (P2:*Selecting an image to interact with a pointing touch-gesture +"That is the Castle XY"*)or switching the order (P3: *"Where is this castle?"+ pointing touch-gesture on the castle.*

For both rounds, the *max-consensus ratio* is medium to high. For the unbiased elicitation, the ratio for caption-related tasks reaches mainly 66.7% to 100%, while the values for the biased elicitation show slightly less consensus with 50% to 100%. Both results indicate that a fairly high number of participants use the combination ST in a similar way. Further, the consensus-distinct ratio shows a wide variety for unbiased (0.33 to 1.00) and biased elicitation (0.25 to 1.00), which indicates a higher variety in multimodal ST interactions.

Mid-air Gesture Participants were not familiar with using mid-air gestures, with 71.4% reporting to have never used it and 28.6% using it less than once per month. Thus, they did not represent a common proposal in the unbiased elicitation. However, proposals increased after applying the priming strategy in the biased elicitation, especially for layout-related tasks. The proposed mid-air gesture interactions are similar to the proposed touch interactions with the pointing gesture or drag and drop to move elements. Also, changing the size of elements

remained similar with a pinched index finger and thumb and moving them diagonally apart or dragging the corner of elements. Longer pointing decreased in proposals, as short pointing was commonly used to select elements and initiate further interaction. Mid-air input was reported to be less strenuous as P9 indicated: "Then I can simply sit here, without needing to go the the screen").

The max-consensus ratio for layout tasks ranges from 50% to 83.3%, showing a slightly lower consensus than the touch interactions. The consensus-distinct ratio for these tasks ranges from 0.25 to 0.50, showing that a wider variety of mid-air interactions is proposed compared to touch interactions (0.50 to 0.75).

Speech and mid-air Gesture Participants were not familiar with mid-air gestures (see mid-air gesture 3.2), but reported high familiarity with speech interactions (see speech 3.2). The combination of speech and gesture (SG) was only proposed in the biased elicitation and resembled the use of ST, with most proposals targeting caption-related tasks. The combination was mostly proposed as an addition by participants who already used ST as a multimodal combination in the unbiased elicitation. The approaches of participants show high similarity to the proposed ST interactions with initiating a verbal command that is then specified via mid-air gesture (P1: *"Add the word castle here instead of farm"+ pointing mid-air-gesture on the castle*).

The max-consensus ratio for caption-related tasks reaches 50% to 100%, showing slightly higher consensus compared to touch interactions. The same tendency is shown in the slightly higher consensus-distinct ratio, which reaches from 0.50 to 1.00, showing that a majority uses mid-air gestures in a similar way.

Eye Tracking Participants were not familiar with using eye tracking, with 71.4% reporting to have never used it and 28.6% using it less than once per month. Eye tracking was only proposed as a common modality in the biased elicitation for the two referents: *1) Insert image* and *6) caption alternative*.

For both referents, the max-consensus ratio is 66.7%, and the consensus-distinct ratio is 0.50, indicating that a majority would use eye-tracking in a similar way. Most participants suggested focusing on elements to initiate further interaction, get alternatives or select elements.

3.3 Ideal Multimodal Photobook Creation

In a questionnaire, we collected how participants imagined an human-AI collaboration. The analysis covers 1) different setups, and 2) how an ideal version of interactive photobook creation would look like.

Setup & Environment Regarding the setup for an interactive photobook co-creation, participants showed a positive tendency towards situated screens, with 12 participants (85.7%) being in favor of them. They highlighted that the larger display helps to *have a good overview* (P12) and creates an *interactive experience* (P2,P7). Only one participant expressed that they would prefer another option *such as a tablet* (P5) for feasibility. Further, a VR setup showed split opinions:

Nine participants (64%) showed enthusiasm to *turn[...] the creation of a photobook into an experience* (P2) and described it as *exciting* (P5,P6,P7,P11) and *fun* (P2,P6,P7). In comparison, five participants are sceptical, expressing that it*doesn't seem to make much sense [...]* to use a 3D environment (P10) or that it is *cumbersome* (P9). Regarding more common setups, a web setup was mostly seen positively by all participants, describing it as a *good* P(5) or even the *easiest* P(10) solution. This aligns with the high level of familiarity that participants recorded in the initial questionnaire. Two participants were more critical, focusing on the advantages of other setups such as *VR* (P4). The mobile application was least preferred, with nine participants (64%) expressing to reject this option. Participants highlighted the *small display* (P2,P4,P5, P6,P7,P8,P9,P10,P11) which leads to *clutter* and a *lack of clarity* (P2,P5,P8,P11).

Co-Creation Process Three different themes were identified for the participants ideal photobook creation: The process should 1) be interactive and multimodal, 2) alternate between automatic and co-creation, and 3) adapt to user preferences.

For the first theme, nine participants (64.3%) expressed that they would like an *interactive tool* (P5) that feels *intuitive* (P3,P6) to use. Further, they would like to use a *combination* of modalities (P4,P4,P5). Most participants highlighted speech as the base and combined it with another modality such as touch, mid-air gestures and eye tracking (P2,P3,P4,P10,P13). They highlighted that the system should support multiple combinations of modalities to be used *easily* and *most intuitively* (P13). However, three participants showed hesitation towards speech and expressed to feel *uncomfortable using it when other people are around* (P14).

Regarding the second theme, eight (57.1%) participants expressed that they would prefer an interaction pattern in which parts are automatically created for them, which they can then adapt in a co-creation process. For example, P7 states *I think it would be quite good if the AI could do a lot of the work for you [...], so that you only have to make minor adjustments* aligning with P8 who describes the process as follows: *I import my photos. A few design questions are asked. Then, a photo book is automatically created that can be customized afterwards.* However, while some automation is preferred, interactive adaptation of the content by the user remains important. P6 states: *Not everything should be done for me, as I also enjoy designing.* While users enjoy needing less effort in engaging with the photobook system, they like to be part of the creation process.

Lastly, five participants (35.7%) expressed that the system should *learn[...] [their] preferences regarding the layout* (P1) and learn *[their] style* P(13). This links to the wish for photobooks to be created more automatically in the previous theme. Users want to *feel understood* (P14) by the system to reduce the effort and *reach [their] goal* faster (P13).

4 Discussion

We conducted an elicitation study to explore the user's perception of commonly used modalities in HCI and to investigate which modalities users find most suit-

able for ML-supported co-creation. Additionally, we investigated how users imagine an ideal IML environment.

Three themes emerged for an ideal ML-assisted co-creation: The process should 1) be interactive and multimodal - to foster an intuitive experience -, 2) alternate between automatic and co-creation - highlighting the need to reduce the workload of the process while maintaining personalization - 3) adapt to user preferences. Aligning with the preference for MMI in theme one, users highlighted situated screens as a suitable setup due its comprehensive overview and enhanced interactivity. Conversely, handheld devices were viewed as least ideal due to their small screen, which resulted in a lack of clarity.

In general, two approaches for human-AI collaboration could be identified throughout the elicitation: Participants either 1) trusted the system's *intelligence* to realize their request, or 2) specified the exact changes the system should realize. Less experience with AI corresponded with less trust into the system to realize their requests. This difference in trust of participants can be linked to two opposing biases in human-AI interaction that are discussed in the literature: While some users experience automation bias - over-reliance in the capabilities of the AI - other users are prone to algorithmic aversions - distrust into the AIs capabilities [45, 52]. To enable a successful interaction, it is necessary to establish a correct-as-possible understanding of the capabilities and limits of the AI system for the user. This can be supported by explanations of the model behavior or additional data in the co-creation process [14, 52]. Future work should thus investigate the impact of XAI methods and additional data on the users understanding of the AI system and their behavior in the co-creation process.

Regarding suitability for ML-assisted co-creation, our findings revealed that speech (S) and touch (T), or a combination of speech and touch (ST), support these tasks best. While the priming in the biased elicitation led to more proposals of uncommon modalities (like mid-air gestures (G)), the most proposed modalities remained the ones that participants were most familiar with [37].

Caption-related tasks were most supported by speech (S) and only showed a low number of proposals for touch (T). This low consideration could be based on the perceived difficulty of transferring language into touch data. Captions represent textual data which is closer to speech and, thus, might be easier for users to verbalize. In comparison, touch (T) or mid-air gestures (G) were mainly applied for layout- and image-related tasks. This could be based on the characteristic of touch to perform actions intuitively with direct manipulation. Transferring spatial information into a verbal expression might be more challenging for the user. This close transfer to spatial data led to layout-related tasks being best supported by touch (T) or mid-air gesture (G) interactions [61]. For image-related tasks, a wide range of modalities is suggested, allowing users to interact intuitively based on their prior experiences. However, the wide range of proposals might point towards problems in finding common modalities.

Further, multimodal suggestions were made in the form of a combination of speech-touch (ST) or speech-gestures (SG), as common for selection tasks [39],. However, the number of multimodal proposals remained low throughout the

elicitation (unbiased:13.4%; biased 14.6%). This contradicts the questionnaire results, which reveal a preference for MMI with theme 1) *The process should be interactive and multimodal.* This finding aligns with indications that providing a multimodal system does not necessarily lead to multimodal interaction [39]. However, the preference for MMI could be due to the high max-consensus and consensus-distinct ratio for the combination of speech and touch (ST). A majority of people would use the combination of speech and touch (ST) in a similar manner when correcting captions, which makes it a suitable modality for ML co-creation. Multimodality was mainly proposed in combination with speech. It specified a verbal command with another modality like touch (T) or mid-air gestures (G). Pointing to an element enables precise corrections, leading to feedback iterations that are faster and contain more information.

Further, it is noticeable that while mid-air gestures (G) were not proposed intuitively in the unbiased elicitation, primed participants proposed mid-air gestures (G), or a combination of speech-gestures (SG) as an alternative to touch (T) or speech and touch (ST) proposals. While enabling multimodal interaction offers higher flexibility, this is only applicable if users are able to make good modality choices [40]. The low experience with gesture interfaces - as indicated in the questionnaire - could prevent the users from making this modality choice, even though it is considered a suitable option when the user is primed. Thus, implementing strategies to help users make better modality choices could help integrate modalities that could be suitable, but are not familiar to the users. One way to address this is the ASPECT and ARCADE Model [26]. ARCADE summarizes six strategies to help people make better modality choices: For example, *A* in ARCADE stands for *Access Information & Experience* and refers to the use of textual, auditive or graphical introductory tutorials. It should be investigated how introducing uncommon modalities with these strategies impacts the user's ability to make good modality choices. Additionally, most users that proposed the multimodal combination speech-touch (ST) in the unbiased round already, proposed the combination speech-gesture (SG) in the biased round. This could indicate that participants who are intuitively prone to use multimodal interactions are also more likely to explore less familiar multimodal interactions.

Eye tracking was only sparsely proposed for two referents in the biased elicitation. The low number of proposals could be tied back to the unfamiliarity of the users with the modality - with 71.4% reporting to never have used it before. Modalities, such as speech and touch, are familiar active modalities that are used consciously, while gaze represents a modality for which users are not used to form active interactions. Based on the foreground-background theory, interactions within a system should combine passive and active interaction [11, 40]. Future research should explore how to combine passive gaze interaction and the modalities that were derived in this work within a foreground-background interaction system. Lastly, it stands out that interactions with controllers were not proposed during the elicitation study. This can be tied to the low amount of experience that participants showed as 42.9% never used controller before.

Implications Based on the evaluation, we derived the following implications: 1) Enabling a wide variety of modalities allows for the most intuitive interactions. 2) Informing users about uncommon modalities opens up suitable modality choices, which are otherwise missed. 3) Multimodal interactions represent a high consensus when chosen by the users. 4) Automation and user-adaptation should be balanced in the creation process.

Limitations & Future Work This paper investigates what modalities users find suitable for which IML task. However, the study was based on elicited proposals. Thus, all findings should be verified on a working prototype in future work. That way, it would be possible to completely compare the techniques, including potential technical limitations [24]. Further, this work indicates how to design interaction for co-creation processes, in which, one relevant aspect is the imperfect performance that the system can display towards the user. This mismatch of system results and user expectations can lead to 1) decreased user trust in the system [41, 31], and 2) adapted user behaviour [40]. For example, an error-prone multimodal system can lead to a shift in the co-timing of user signals to clarify their behaviour to the system (e.g., hypertiming) [40]. Thus, future work should investigate the impact of imperfect performance in co-creation processes on trust and user behaviour, as well as explore strategies to handle model ambiguity in human-AI collaboration. One possible strategy is the use of Explainable Artificial Intelligence (XAI) to give insight into the model behaviour [28].

5 Conclusion

Our paper explored user's perceptions on modalities in HCI in the context of co-creation for ML tasks. Photobook creation was applied as a use case as it includes a variety if ML tasks like image-, caption-, and layout-related tasks. We suggested a list of 14 common operations for ML tasks and conducted an elicitation study with 14 participants. Additionally, through a questionnaire, we investigated how users imagine an ideal ML co-creation environment. Our findings indicate that 1) interaction and multimodality, 2) balance between automation and co-creation, and 3) adaptation to user preferences, represent relevant themes for co-creation. Even though multimodal proposals remained sparse in the elicitation, the participant's proposals of speech combined with touch or mid-air gestures showed high consensus between participants, matching the highlighted preference of MMI. The modalities speech and touch - either individually or combined, effectively support various tasks in ML co-creation. Caption-related tasks are best handled by speech, layout-related tasks are most effectively managed through touch or mid-air gestures, and image-related tasks, are supported by a variety of modalities. Thus, providing a system that covers multiple modalities enables users to interact intuitively based on their prior experiences. Finally, introducing uncommon modalities - such as mid-air gestures - to the user can open up suitable modality choices, that are otherwise missed.

Acknowledgments. This work was funded by the German Federal Ministry of Education and Research (BMBF) under grant numbers 01IW23002 (No-IDLE) and 01IW24006 (NoIDLEChatGPT), as well as by the Endowed Chair of Applied AI at the University of Oldenburg.

References

1. Amershi, S., Cakmak, M., Knox, W.B., Kulesza, T.: Power to the People: The Role of Humans in Interactive Machine Learning. AI Magazine **35**(4), 105–120 (Dec 2014). https://doi.org/10.1609/aimag.v35i4.2513, https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2513, number: 4
2. Anagnostopoulou, A., Hartmann, M., Sonntag, D.: Putting humans in the image captioning loop. arXiv preprint arXiv:2306.03476 (2023)
3. Aurisano, J., Kumar, A., Gonzales, A., Reda, K., Leigh, J., Di Eugenio, B., Johnson, A.: Show me data": Observational study of a conversational interface in visual data exploration. In: IEEE VIS. vol. 15, p. 1 (2015)
4. Barnaby, C., Chen, Q., Wang, C., Dillig, I.: Photoscout: Synthesis-powered multimodal image search. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. pp. 1–15 (2024)
5. Barz, M., Sonntag, D.: Automatic visual attention detection for mobile eye tracking using pre-trained computer vision models and human gaze. Sensors **21**(12), 4143 (2021)
6. Beşevli, C., Buruk, O.T., Erkaya, M., Özcan, O.: Investigating the effects of legacy bias: User elicited gestures from the end users perspective. In: Proceedings of the 2018 ACM Conference Companion Publication on Designing Interactive Systems. pp. 277–281 (2018)
7. Bossen, C., Dindler, C., Iversen, O.S.: Evaluation in participatory design: a literature survey. In: Proceedings of the 14th Participatory Design Conference: Full papers-Volume 1. pp. 151–160 (2016)
8. Bown, O., Brown, A.R.: Interaction design for metacreative systems. New Directions in Third Wave Human-Computer Interaction: Volume 1-Technologies **1**, 67–87 (2018)
9. Braun, V., Clarke, V.: Thematic analysis. American Psychological Association (2012)
10. Braun, V., Clarke, V.: Reflecting on reflexive thematic analysis. Qualitative research in sport, exercise and health **11**(4), 589–597 (2019)
11. Buxton, B.: Integrating the periphery & context: A new model of telematics. In: Graphics Interface. pp. 239–239. Canadian Information Processing Society (1995)
12. Choudhary, K., DeCost, B., Chen, C., Jain, A., Tavazza, F., Cohn, R., Park, C.W., Choudhary, A., Agrawal, A., Billinge, S.J., et al.: Recent advances and applications of deep learning methods in materials science. npj Computational Materials **8**(1), 59 (2022)
13. Cox, K., Grinter, R.E., Hibino, S.L., Jagadeesan, L.J., Mantilla, D.: A multi-modal natural language interface to an information visualization environment. International Journal of Speech Technology **4**, 297–314 (2001)
14. De-Arteaga, M., Fogliato, R., Chouldechova, A.: A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In: Proceedings of the 2020 CHI conference on human factors in computing systems. pp. 1–12 (2020)

15. Duchowski, A.T.: Gaze-based interaction: A 30 year retrospective. Computers & Graphics **73**, 59–69 (Jun 2018). https://doi.org/10.1016/j.cag.2018.04.002, https://www.sciencedirect.com/science/article/pii/S0097849318300487
16. Dudley, J.J., Kristensson, P.O.: A Review of User Interface Design for Interactive Machine Learning. ACM Transactions on Interactive Intelligent Systems **8**(2), 1–37 (Jun 2018). https://doi.org/10.1145/3185517, https://dl.acm.org/doi/10.1145/3185517
17. Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G., Beck, H.P.: The role of trust in automation reliance. International journal of human-computer studies **58**(6), 697–718 (2003)
18. Fails, J.A., Olsen Jr, D.R.: Interactive machine learning. In: Proceedings of the 8th international conference on Intelligent user interfaces. pp. 39–45 (2003)
19. Gmeiner, F., Holstein, K., Martelaro, N.: Team learning as a lens for designing human-ai co-creative systems (2022). https://doi.org/10.48550/arXiv.2207.02996, https://arxiv.org/abs/2207.02996
20. Good, M.D., Whiteside, J.A., Wixon, D.R., Jones, S.J.: Building a user-derived interface. Communications of the ACM **27**(10), 1032–1043 (1984)
21. Griffin, Z.M., Bock, K.: What the eyes say about speaking. Psychological science **11**(4), 274–279 (2000)
22. He, Z., Li, S., Song, Y., Cai, Z.: Towards building condition-based cross-modality intention-aware human-ai cooperation under vr environment. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. pp. 1–13 (2024)
23. Henze, N., Boll, S.: Who's that girl? handheld augmented reality for printed photo books. In: Human-Computer Interaction–INTERACT 2011: 13th IFIP TC 13 International Conference, Lisbon, Portugal, September 5-9, 2011, Proceedings, Part III 13. pp. 134–151. Springer (2011)
24. Herbig, N., Pal, S., Van Genabith, J., Krüger, A.: Multi-Modal Approaches for Post-Editing Machine Translation. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. pp. 1–11. ACM, Glasgow Scotland Uk (May 2019). https://doi.org/10.1145/3290605.3300461, https://dl.acm.org/doi/10.1145/3290605.3300461
25. Jacob, R.J.K.: What you look at is what you get: eye movement-based interaction techniques. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 11–18. CHI '90, Association for Computing Machinery, New York, NY, USA (Mar 1990). https://doi.org/10.1145/97243.97246, https://dl.acm.org/doi/10.1145/97243.97246
26. Jameson, A., Berendt, B., Gabrielli, S., Cena, F., Gena, C., Vernero, F., Reinecke, K., et al.: Choice architecture for human-computer interaction. Foundations and Trends® in Human–Computer Interaction **7**(1–2), 1–235 (2014)
27. Jameson, A.D., Spaulding, A., Yorke-Smith, N.: Introduction to the special issue on "usable ai". AI Magazine **30**(4), 11–11 (2009)
28. Kadir, M.A., Mosavi, A., Sonntag, D.: Evaluation Metrics for XAI: A Review, Taxonomy, and Practical Applications. In: 2023 IEEE 27th International Conference on Intelligent Engineering Systems (INES). pp. 000111–000124. IEEE (2023), https://ieeexplore.ieee.org/abstract/document/10297629/
29. Kantosalo, A., Ravikumar, P.T., Grace, K., Takala, T.: Modalities, styles and strategies: An interaction framework for human-computer co-creativity. In: ICCC. pp. 57–64. International Conference on Computational Creativity, Online (2020)
30. Kath, H., Gouvêa, T.S., Sonntag, D.: A human-in-the-loop tool for annotating passive acoustic monitoring datasets. In: IJCAI. pp. 7140–7144 (2023)

31. Kocielnik, R., Amershi, S., Bennett, P.N.: Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. pp. 1–14 (2019)
32. Luo, Y., Yu, J., Liang, M., Wan, Y., Zhu, K., Santosa, S.S.: Emotion embodied: Unveiling the expressive potential of single-hand gestures. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. pp. 1–17 (2024)
33. Maszuhn, M., Abdenebaoui, L., Boll, S.: A user-centered approach for recognizing convenience images in personal photo collections. In: 2021 International Conference on Content-Based Multimedia Indexing (CBMI). pp. 1–4. IEEE (2021)
34. Maybury, M., Wahlster, W.: Readings in intelligent user interfaces. Morgan Kaufmann (1998)
35. McGuire, J., De Cremer, D., Van de Cruys, T.: Establishing the importance of co-creation and self-efficacy in creative collaboration with artificial intelligence. Scientific Reports **14**(1), 18525 (Aug 2024). https://doi.org/10.1038/s41598-024-69423-2, https://www.nature.com/articles/s41598-024-69423-2, publisher: Nature Publishing Group
36. Morris, M.R.: Web on the wall: insights from a multimodal interaction elicitation study. In: Proceedings of the 2012 ACM international conference on Interactive tabletops and surfaces. pp. 95–104. ACM, Cambridge Massachusetts USA (Nov 2012). https://doi.org/10.1145/2396636.2396651, https://dl.acm.org/doi/10.1145/2396636.2396651
37. Morris, M.R., Danielescu, A., Drucker, S., Fisher, D., Lee, B., schraefel, m.c., Wobbrock, J.O.: Reducing legacy bias in gesture elicitation studies. Interactions **21**(3), 40–45 (May 2014). https://doi.org/10.1145/2591689, https://doi.org/10.1145/2591689
38. Nielsen, M., Storring, M., Moeslund, T.B., Granum, E.: A procedure for developing intuitive and ergonomic gesture interfaces for man-machine interaction. Aalborg, Denmark (2003)
39. Oviatt, S.: Ten myths of multimodal interaction. Communications of the ACM **42**(11), 74–81 (1999)
40. Oviatt, S., Schuller, B., Cohen, P.R., Sonntag, D., Potamianos, G., Krüger, A. (eds.): The Handbook of Multimodal-Multisensor Interfaces: Foundations, User Modeling, and Common Modality Combinations - Volume 1, vol. 14. Association for Computing Machinery and Morgan & Claypool (Mar 2017). https://doi.org/10.1145/3015783
41. Papenmeier, A., Kern, D., Englebienne, G., Seifert, C.: It's complicated: The relationship between user trust, model accuracy and explanations in ai. ACM Transactions on Computer-Human Interaction (TOCHI) **29**(4), 1–33 (2022)
42. Rezwana, J., Maher, M.L.: Designing Creative AI Partners with COFI: A Framework for Modeling Interaction in Human-AI Co-Creative Systems. ACM Transactions on Computer-Human Interaction **30**(5), 1–28 (Oct 2023). https://doi.org/10.1145/3519026, https://dl.acm.org/doi/10.1145/3519026
43. van Rijn, P., Mertes, S., Janowski, K., Weitz, K., Jacoby, N., André, E.: Giving robots a voice: Human-in-the-loop voice creation and open-ended labeling. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. pp. 1–34 (2024)
44. Sandhaus, P., Thieme, S., Boll, S.: Processes of photo book production. Multimedia Systems **14**, 351–357 (2008)

45. Schecter, A., Bogert, E., Lauharatanahirun, N.: Algorithmic appreciation or aversion? the moderating effects of uncertainty on algorithmic decision making. In: Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems. pp. 1–8 (2023)
46. Schuler, D., Namioka, A.: Participatory Design: Principles and Practices. CRC Press (Mar 1993), google-Books-ID: pWOEk6Sk4YkC
47. Sonntag, D., Barz, M., Gouvêa, T.: A look under the hood of the interactive deep learning enterprise (no-idle) (6 2024), https://arxiv.org/abs/2406.19054
48. Sonntag, D., Gouvea, T., Barz, M., Anagnostopoulou, A., Liang, S., Bittner, S.J., Scheurer, F.: The Interactive Deep Learning Enterprise (No-IDLE) meets ChatGPT. Tech. rep., German Research Center for AI (2024)
49. Talamo, M.: The Digital Revolution and the Art of Co-creation. In: Arbizzani, E., Cangelli, E., Clemente, C., Cumo, F., Giofrè, F., Giovenale, A.M., Palme, M., Paris, S. (eds.) Technological Imagination in the Green and Digital Transition. pp. 27–35. Springer International Publishing, Cham (2023). https://doi.org/10.1007/978-3-031-29515-7_4
50. Turk, M.: Multimodal interaction: A review. Pattern recognition letters **36**, 189–195 (2014)
51. Tusfiqur, H.M., Nguyen, D.M., Truong, M.T., Nguyen, T.A., Nguyen, B.T., Barz, M., Profitlich, H.J., Than, N.T., Le, N., Xie, P., et al.: Drg-net: interactive joint learning of multi-lesion segmentation and classification for diabetic retinopathy grading. arXiv preprint arXiv:2212.14615 (2022)
52. Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M.S., Krishna, R.: Explanations can reduce overreliance on ai systems during decision-making. Proceedings of the ACM on Human-Computer Interaction **7**(CSCW1), 1–38 (2023)
53. Villarreal-Narvaez, S., Vanderdonckt, J., Vatavu, R.D., Wobbrock, J.O.: A systematic review of gesture elicitation studies: What can we learn from 216 studies? In: Proceedings of the 2020 ACM designing interactive systems conference. pp. 855–872 (2020)
54. Wang, Z., Huang, Y., Song, D., Ma, L., Zhang, T.: Promptcharm: Text-to-image generation through multi-modal prompting and refinement. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. pp. 1–21 (2024)
55. Wobbrock, J.O., Aung, H.H., Rothrock, B., Myers, B.A.: Maximizing the guessability of symbolic input. In: CHI '05 Extended Abstracts on Human Factors in Computing Systems. pp. 1869–1872. CHI EA '05, Association for Computing Machinery, New York, NY, USA (Apr 2005). https://doi.org/10.1145/1056808.1057043, https://doi.org/10.1145/1056808.1057043
56. Wobbrock, J.O., Morris, M.R., Wilson, A.D.: User-defined gestures for surface computing. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 1083–1092. CHI '09, Association for Computing Machinery, New York, NY, USA (Apr 2009). https://doi.org/10.1145/1518701.1518866, https://doi.org/10.1145/1518701.1518866
57. Wu, Z., Ji, D., Yu, K., Zeng, X., Wu, D., Shidujaman, M.: AI Creativity and the Human-AI Co-creation Model. In: Kurosu, M. (ed.) Human-Computer Interaction. Theory, Methods and Tools. pp. 171–190. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-78462-1_13
58. Yang, J., Shi, Y., Zhang, Y., Li, K., Rosli, D.W., Jain, A., Zhang, S., Li, T., Landay, J.A., Lam, M.S.: Reactgenie: A development framework for complex multimodal interactions using large language models. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. pp. 1–23 (2024)

59. Yasen, M., Jusoh, S.: A systematic review on hand gesture recognition techniques, challenges and applications. PeerJ Computer Science **5**, e218 (2019)
60. Yuksel, B.F., Kim, S.J., Jin, S.J., Lee, J.J., Fazli, P., Mathur, U., Bisht, V., Yoon, I., Siu, Y.T., Miele, J.A.: Increasing video accessibility for visually impaired users with human-in-the-loop machine learning. In: Extended abstracts of the 2020 CHI conference on human factors in computing systems. pp. 1–9 (2020)
61. Zhou, X., Williams, A.S., Ortega, F.R.: Eliciting multimodal gesture+ speech interactions in a multi-object augmented reality environment. In: Proceedings of the 28th ACM Symposium on Virtual Reality Software and Technology. pp. 1–10 (2022)

A Appendix

A.1 Briefing of Study Context

You would like to create a travel photo book from your last trip. To do this, you use a new system for creating a photo book digitally. The system is supported by AI and gives you the freedom to choose how you want to interact with it. All technologies are possible and whatever you have in mind - the photo book creation system will understand you and execute your input as intended. A photo book is created from your input. The study is a prototype that simulates the functionalities you will use. The prototype is kept simple by intention to give you an idea of the user interface but enable you to include own ideas and creativity.

A.2 Priming Strategy - List of Modalities

The following table displays the list of five modalities with corresponding descriptions which were shown to participants in a step of the priming strategy for the user study. After reading the content of this table participants first watched videos for each referent in which the interaction with the modality was shown. Then they were asked to propose interactions for an example referent to ensure that the participants understood the modality sufficiently.

A.3 Questionnaire

The following table displays the questions of the final questionnaire. It contains 13 questions, divided into six categories: Current digital photobook creation, ideal photobook creation, ideal photobook creation setup, familiarity with modalities, experience with photobook creation and additional questions. It consists of open-ended questions and 5-point Likert scales.

Final Questionnaire – Ideal Interactive Photobook Creation	
Question	**Type**
Current digital photobook creation	
1. What functionalities are you using now in digital photobook creation? 2. What did you like about creating a digital photobook? 3. What did you dislike about creating a digital photobook? 4. Based on your experience with the creation of digital photo books, what functions are you missing?	Open-ended question
Ideal photobook creation	
5. Describe an ideal AI-assisted photobook creation.	Open-ended question
Ideal photobook creation setup	
6. What do you think about creating a photobook on a situated screen (similar to the test scenario)? 7. What do you think about creating a photobook in virtual reality? 8. What do you think about creating a photobook in a mobile app? 9. What do you think about creating a photobook in a web/desktop app?	Open-ended question
Familiarity with modalities	
10. How often do you use the following technologies? *(never, less than once per month, between once a week and once a month, a few times a week, almost every day)* a. Computer/Laptop b. Virtual Reality (VR) c. Gesture Recognition d. Speech Input e. Gaze-Tracking f. Controller (VR, Xbox, PlayStation, Wii)	5-point Likert scale
Experience with photobook creation	
11. On a scale of 1 to 5, how would you rate your experience with digital photobook creation? *(1 = very inexperienced, 5 = very experienced)* 12. On a scale of 1 to 5, how complicated did you find creating digital photo books? *(1 = very simple, 5 = very complicated)*	5-point Likert scale
In addition	
13. Is there anything you would like to add?	Open-ended question

Fig. 2. Content of the final questionnaire including questions about ideal photobook creation

Table 4. List of modalities and corresponding explanations for the priming strategy

Modality	Description
Speech	The system will pick up your speech and understand your intended action. The system will then implement your command.
Eye Tracking	The system can monitor what you are looking at. For one, the system knows what you are referring to on the screen (photobook page, images, captions, chapter overview etc.). For another, it can interpret intended eye movements such as blinks or focusing on an element of the photobook.
Controller	The system will pick up where the controller is pointing at and you can interact with the systems by actions such as click, circle, drag or keep pressed, shake and more.
Mid-Air Gestures	The system will pick up your mid-air hand movements and understand your intended action. The system will then implement your command.
Touch	The system will pick up your hand movements on the screen and understand your intended action. The system will then implement your command.

A.4 Results

Table 5. Unbiased Elicitation: Proposals per referent (Ref) in the unbiased elicitation study - Extension. The max-consensus (Cm) and consensus-distinct (Cd , threshold = 2) ratios for all and the modalities suggested ≥ 3 times (S=Speech, T=Touch, XY=combination of X and Y, e.g. ST for speech and touch).

Nr	Referent	All		S		T		ST	
		C_m	Cd	C_m	C_d	C_m	C_d	C_m	C_d
1	insert image	57.1	0.80	75	0.50	87.5	0.00	66.7	0.50
2	image selection	64.3	0.60	75	0.33	55.6	0.66	-	-
3	image alternatives	57.1	0.50	87.5	0.50	50	0.33	-	-
4	image search	71.4	0.12	80	1.00	-	-	-	-
5	similar images	64.3	0.80	88.9	0.66	-	-	-	-
6	captions alternative	42.9	0.57	100	1.00	-	-	100	1.00
7	caption style	57.1	0.75	100	1.00	66.6	1.00	-	-
8	caption context	78.6	0.4	63.6	0.75	-	-	66.7	0.50
9	caption location	64.3	0.6	88.9	1.00	-	-	66.7	0.50
10	caption person	64.3	0.6	88.9	0.50	-	-	75.0	0.33
11	caption correction	57.1	0.33	100	1.00	-	-	66.7	0.50
12	move elements	64.3	0.60	66.7	0.50	87.5	0.50	-	-
13	change image size	64.2	0.33	-	-	55.6	0.66	-	-
14	change font size	44.4	0.75	-	-	44.4	0.75	-	-

Table 6. Biased Elicitation: Proposals per referent (Ref) in the biased elicitation study - Extension. The max-consensus (Cm) and consensus-distinct (Cd , threshold = 2) ratios for all and the modalities suggested ≥ 3 times (S=Speech, T=Touch, G=Gesture, E=Eye, C= Controller, XY=combination of X and Y, e.g. ST for speech and touch).

Nr	Referent	All		S		T		G		ST		E		SG	
		C_m	C_d	C_m	C_d	C_m	C_d	C_m	C_d	C_m	C_d	C_m	C_d	C_m	C_d
1	insert image	78.6	1.0	75	0.50	81.8	0.0	-	-	75	0.50	66	0.5	-	-
2	image selection	78.6	0.80	75	0.33	54.5	0.66	66.7	0.33	-	-	-	-	-	-
3	image alternatives	64.3	0.5	77.8	1	60	0.33	-	-	-	-	-	-	-	-
4	image search	78.6	0.29	81.8	1	-	-	-	-	50	0.25	-	-	-	-
5	similar images	64.3	0.80	88.9	0.67	60	1	-	-	66.7	0.5	-	-	-	-
6	captions alternative	42.9	0.83	100	1.0	50	1.00	-	-	100	1	66.7	0.5	100	1.00
7	caption style	57.1	1	100	1	75	1	-	-	80	0.5	-	-	-	-
8	caption context	78.6	0.60	63.6	0.75	-	-	-	-	75	0.5	-	-	-	-
9	caption location	64.3	0.60	88.9	1.0	-	-	-	-	50	0	-	-	75	0.50
10	caption person	64.3	0.80	88.9	0.5	-	-	-	-	60	0.25	-	-	75	0.50
11	caption correction	78.6	0.50	100	1.00	-	-	-	-	50	0.3	-	-	66.7	0.5
12	move elements	78.6	0.83	66.7	0.5	90.1	0.50	50	0.25	-	-	-	-	-	-
13	change image size	71.4	0.42	-	-	60	0.67	83.3	0.50	-	-	-	-	-	-
14	change font size	71.4	0.0	-	-	50	0.75	80	0.5	-	-	-	-	-	-

A.5 Analysis

Table 7. The 6 stages of the Reflexive Thematic Analysis by Braun et al.(2019) [10]

Stage	Explanation
Familiarize	The researcher familiarize themselves with the data set. The questionnaire answers are read through several times. Additionally, notes are taken.
Coding	The data set is coded throughout several iterations to break the data into manageable content.
Generate Initial Themes	Patterns are explored through the re-grouping of codes. These form potential themes.
Reviewing and Developing Themes	The initial themes are checked if they hold the true meaning of the data set and then are further developed. This can include the splitting or fusion of themes.
Refining, Defining, and, Naming Themes	Themes are named. Moreover, the scope and concept of the themes are defined.
Producing the Report	The story of the data set is presented. This involved the themes, the definition of the themes, and examples of the data set as well as their interpretation.