2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR) | 979-8-3503-7402-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/VR58804.2024.00034

OdorAgent: Generate Odor Sequences for Movies Based on Large Language Model

Yu Zhang*
Academy of Arts & Design, Tsinghua University

Peizhong Gao †
Tsinghua University

Fangzhou Kang ‡
Xinya College, Tsinghua University

Jiaxiang Li §
Academy of Arts & Design, Tsinghua University

Jiacheng Liu ¶
Xinya College, Tsinghua University

Qi Lu ‖
The Future Laboratory, Tsinghua University

Yingqing Xu **
The Future Laboratory, Tsinghua University
Academy of Arts & Design, Tsinghua University

Abstract

Numerous studies have shown that integrating scents into movies enhances viewer engagement and immersion. However, creating such olfactory experiences often requires professional perfumers to match scents, limiting their widespread use. To address this, we propose OdorAgent which combines a LLM with a text-image model to automate video-odor matching. The generation framework is in four dimensions: subject matter, emotion, space, and time. We applied it to a specific movie and conducted user studies to evaluate and compare the effectiveness of different system elements. The results indicate that OdorAgent possesses significant scene adaptability and enables inexperienced individuals to design odor experiences for video and images.

Index Terms: Human-centered computing—Interaction design—Systems and tools for interaction design—

1 Introduction

Although visual and auditory elements capture the majority of people's attention, there has been a persistent endeavor to amalgamate additional sensory experiences to create more immersive and captivating content [17,33,4]. Olfactory is intimately tied to human memory and emotion [14,46], with extensive commercial applications. The incorporation of scents into movies has a long history, and with the evolution of Virtual Reality (VR), olfactory display devices have found their place, enhancing people's gaming or video-watching experiences [29,30,6,31]. In the academic realm, many studies have demonstrated that combining the visual senses with olfactory can intensify immersion [4,26,38]. Nevertheless, the olfactory experience design in multisensory experiences needs specialized experts for olfactory design, with varying styles among different experts, constraining the widespread application of olfactory experiences. To meet the burgeoning demand for multisensory experiences, a digital method for generating scent sequences is required to match a large number of audiovisual entertainment resources, such as games and movies.

*e-mail: yu-zhang21@mails.tsinghua.edu.cn
†e-mail: gpz23@mails.tsinghua.edu.cn
‡e-mail: kfz19@mails.tsinghua.edu.cn
§e-mail: lijiaxia21@mails.tsinghua.edu.cn
¶e-mail: jc-liu21@mails.tsinghua.edu.cn
‖e-mail: luq@mail.tsinghua.edu.cn, corresponding author
**e-mail: yqxu@tsinghua.edu.cn, corresponding author

The design of odor sequences that matching images presents is a complex question, while the emergence of large-scale models (LLM) offers new possibilities for addressing issues in multisensory olfactory experience design. As the use of textual labels to describe scents remains the prevailing method for describing olfactory experiences, it is possible to use the rich corpus of large models to design odor matching for key images. Nonetheless, designing odor delivery for audiovisual entertainment resources still poses a diverse set of challenges. Unlike vision, which directly provides abundant information, olfactory operates indirectly, enhancing other sensory functions, evoking memories, and stimulating associations. The mechanism of this effect may vary among individuals. Furthermore, the smell does not have the mixing pattern like color, the implementation of odor delivery typically needs pre-prepared odor sources, the quality of olfactory experiences is thus restricted by the pre-prepared scent materials.

This paper explores this issue, drawing upon research in past olfactory experience design and expert interviews. In this paper, we present the OdorAgent olfactory experience design framework, capable of utilizing large language models to generate scent sequences aligned with videos. The framework consists of two components: image informations extraction and scent selection methods. We employ computer vision models to extract foreground, background, emotional, and foreground-background distance information from the visual scene. Subsequently, based on these informations, large language models select appropriate scents from a scent library. To enable the precise selection of scents by large language models, we categorize olfactory features into three types: 1. foreground scents representing the main subject of the scene, 2. background scents (including environmental and emotional background scents), and 3. the blending ratios of different scents.

The goal of OdorAgent is to offer a universal olfactory experience design framework that does not require human experts' involvement. As no similar work has been done before, there are some unclear details that needs further exploration and discussion: Can olfactory experiences designed using this framework enhance user immersion? Will the blending of multiple scents be off-putting to users? To what extent do the three categories of olfactory features introduced by OdorAgent affect olfactory experiences? How does OdorAgent perform in different types of scenarios? To address these questions, we conducted user experiments. Taking a movie as an example, we selected a movie segment containing various scenes with different olfactory features and invited 16 volunteers to participate in the experiment. By having each volunteer watch the same movie segment with different olfactory features, we systematically evaluate the influence of three olfactory features we proposed in olfactory experience design.

In summary, this paper's main contributions are as follows:

979-8-3503-7402-5/24/$31.00 ©2024 IEEE
DOI 10.1109/VR58804.2024.00034

- The OdorAgent olfactory experience design framework, which proposes how to use large language models for olfactory experience design in place of olfactory design experts.
- Conducting user experiments to investigate the effectiveness of applying large models to olfactory experience design, along with an in-depth analysis of the effects of different olfactory features within the framework.
- Obtaining user feedback on a universal olfactory design framework to provide insights and references for future work.

2 Related Work

2.1 Digital Scent Technology in Immersive Environment

The development of digital scent technology has brought olfactory stimuli as a part of mulsemedia to immersive environments, especially Virtual Reality. One of the earliest studies of the olfactory experience applied in immersive environments is Sensorama Simulator [13], a multisensory VR system to improve the viewing experience and enhance the sense of presence. The system allows users to sense the wind, vibration, and smell simultaneously. Since then, similar works like Birdly [39] and Advanced Virtual Environment Real-Time Fire Trainer (ADVERT-FT) [8] were created to continue the study of the coordination of odor with other stimuli. More recently, Ranasinghe et al. [38] proposed a system, Season Traveller, that provides users with novel experiences to sense the changes of the four seasons in VR through the joint action of olfactory, wind and thermal simulation. In the paper, the pilot studies are described in detail, including people's perceptions of odors and seasons, and the selection of odor display techniques. Research has revealed that digital scent technologies play a positive role on immersive experience in enhancing the sense of presence that contributes to positive emotions and enjoyment [3,31,28,24,23,22]. However, these studies focuse on the overall effect and evaluation of multisensory system rather than on individual olfactory experiences.

2.2 Movies with scents

In the last century, people began experimenting with introducing scents during movie or performance viewing. For instance, Rothafel used electric fans to waft the aroma of rose-scented cotton towards the audience, allowing them to watch the performance amidst a fragrant atmosphere [16, 37]. Subsequently, there was a gradual shift towards utilizing scents to enhance storytelling. In 1960, the film "Scent of Mystery," starring Elizabeth Taylor, introduced scent-enhanced cinema experiences by modifying seats to release scents related to clues about the identity of the perpetrator [11].

The pursuit of sensory entertainment has motivated people to continuously attempt to combine the sense of smell with visual experiences. However, due to the unique nature of scents, achieving this has not been straightforward. Not all individuals appreciate scent-enhanced movies due to issues such as noise generated during scent release, disliked scents, or delays in updating scents from previous scenes. With advancements in technology, various scent delivery techniques have emerged. These include SensaBubble [41], which employs bubbles to transmit scents; air cannons for scent delivery; ink jet olfactory display [42]; ultrasonic atomization for scent diffusion; and scent cards that release odors through scratching [15], etc.

Odor-enhanced films have not been commercially successful largely because there are great challenges in applying digital scent technology to movies. The simulation of olfaction is limited by a fundamental scientific problem — the smell cannot be arbitrarily synthesized as opposed to the light. There are no components of smell as condensed as the three primary colors of light that can simulate a broad range of visual effects. Consequently, the range of olfactory stimuli that can be effectively conveyed is inherently limited [9, 35, 34, 19, 2, 27, 12, 10].

As a result, odor display technology lags far behind visual display technology. A prosperous future of smell has been depicted but the pathway is still unclear. For instance, The alignment of scents with visuals typically involves manually annotating scent sequences for scent-enhanced movies, predefining scent content, and triggering scent release using specific commands [6].

2.3 Odor Display Based on Image

With the evolution of intelligent methods, some researchers have introduced screen-based intelligent scent generation approaches. Some have trained a CNN network using supervised methods, which outperformed typical CNN networks. This network can simultaneously identify the location and size of scent sources. Additionally, a scent dispersal device connected to a smartphone is used to recognize objects captured by the smartphone's camera and release corresponding scents [18].

Similarly, the OlfacEnhancer also identifies objects through a camera and emits corresponding scents. This portable device can evoke olfactory sensations even when there is some distance between the individual and the object [20]. Researchers like Safaa Alraddadi have processed both visual and audio information to trigger scents automatically using deep learning techniques. They believe this approach can be applied to scenarios such as movies and games but requires more data for support [1].

Georgios Tsaramirsis and others have developed a Unity3D game development resource. When users collide with game objects in the virtual world, it activates the corresponding scent [3]. Presently, efforts to generate scents based on visuals primarily focus on identifying objects in the visuals and subsequently releasing the corresponding pre-prepared scents [43].

2.4 Large Language Model Based Agents and Prompt Engineering

The rapid development of Large Language Models (LLM), such as GPT-4, has resulted in excitement among both technical personnel and designers, who are increasingly leveraging these models to support their work. Notably, one area that has witnessed significant progress is the study of Agents. With the aid of LLM, agents can now execute complex decisions and perform tasks that were previously exclusive to human beings [5,7,47,50]. Park et al. proposed Generative Agents—computational software agents that simulate believable human behavior, and performed simulation in an interactive sandbox [36]. Similarly, there is work exploring agent empowerment in open games such as Minecraft [44,32]. Prompt engineering methods such as Chain of Thought and ReAct were developed to enhance the controllability and accuracy of LLM in task completion, while also endowing it with new abilities. These methods have been instrumental in improving the performance of LLM and enabling it to carry out more complex tasks. Wei et al. described Chain-of-thought, "a series of intermediate reasoning steps", increases the capacity of large language models to carry out sophisticated reasoning [45]. While ReAct, using a simple Wikipedia API, avoids common issues like hallucination and error propagation found in chain-of-thought reasoning. It produces task-solving paths resembling human behavior more than baselines without reasoning traces [49]. Similarly, MM React was proposed to promote ChatGPT for multimodal reasoning and action. It allows ChatGPT to seek help from visual experts (other visual models) and search engines to analyze the content of images and link them to real-world information [48]. To date, no research has been conducted to investigate the potential of utilizing large model-based agents in the domain of odor perception.

3 OdorAgent System

The OdorAgent should be capable of using large language models to generate scent sequences aligned with videos. Numerous researchers have also delved into the field of olfactory experience design [21, 26,

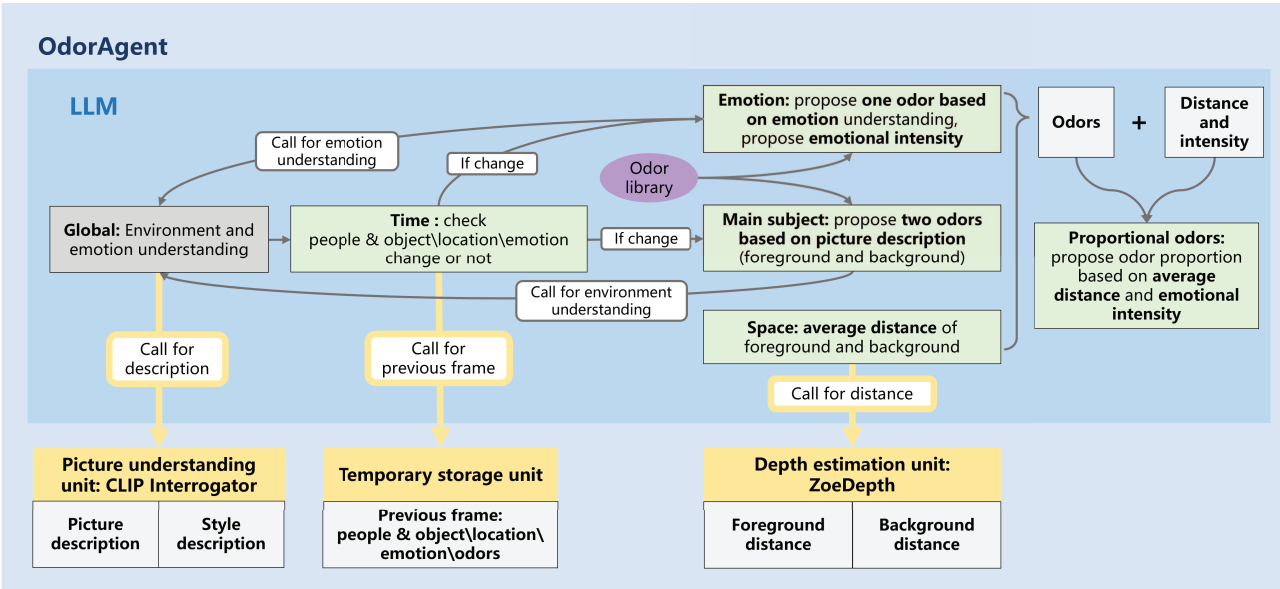

Figure 1: OdorAgent sequentially invokes different models to obtain information related to the picture's odor and uses LLM to gradually understand and process this information and give the odors that match this picture.

25]. According to their research findings, four dimensions need to be taken into account in olfactory experience design : identifying what objects are, evoking various emotions, judging the distance from the source of the odor, and odor duration. To enable large language models to conduct rational olfactory experience design, we conducted interviews with two experts in the field of olfactory design. Drawing from their knowledge and experience, they provided the following recommendations for OdorAgent:

- Swift transitions between scents should take into account the diffusion of odors, ensuring that previously played scents do not affect subsequently played ones.
- Whenever possible, it is advisable to use authentic scents, as genuine scents evoke a more natural response in individuals.
- Olfactory is a highly personal experience intimately connected to individual memories, and different individuals may perceive the same scent differently.
- When designing, it is essential to consider emotional characteristics.

We have integrated insights from olfactory design experts and previous research to enable OdorAgent to select the appropriate scents based on four dimensions of information: the presence of people and objects in the visual scene, the emotions conveyed by the plot, the spatial location of odor sources, and the detection of scene transitions. By selecting scents based on the presence of people and objects in the visual scene, we ensure that the choice of scents aligns with human perception, minimizing cognitive conflicts. Choosing scents that align with the emotional experiences conveyed in the visual scene deepens viewers' memories and immersion, while avoiding scents that contradict those emotions. When odor sources are far, the scent's concentration should be lower, consistent with human perception. Typically, the main elements of a scene are situated in the foreground, representing the focal point of the image. Consequently, the concentration of scents in this portion of the scene should be higher than that of background scents. Detecting scene transitions allows for the timely cessation of scent delivery that does not match the ongoing visual content.

The sensory experience of interaction is influenced by both the authenticity of odors and the compatibility of odors with the video. Since odors cannot be generated out of nothing, whether indicating objects appearing in the frame or evoking different emotions, olfactory experience relies on odor sources. OdorAgent can establish a connection with real odors in two ways: relying on the knowledge of large language models, allowing large language models to associate possible odors and then search for corresponding odors, or pre-informing large language models about a prepared odor library, requiring large language models to select odors from this library based on its understanding of odor labels.

We have implemented this concept using ChatGPT (gpt-3.5-turbo-instruct). To better design olfactory content for movie scenes, ChatGPT needs to first comprehend all the information in the frame. Then, by providing appropriate prompts, ChatGPT can generate the desired results, as shown in Fig. 1.

- Global: When an image is input to ChatGPT, it first needs to understand all the information in the frame. At this point, ChatGPT invokes the Clip Interrogator to obtain a description of the image's content and its stylistic description. Through suitable prompts, ChatGPT provides information about the individuals, objects, and environmental context in the image based on the content description and a sentence regarding the emotional characteristics conveyed by the image's content description and style description.
- Time: When the image changes, the olfactory sequences should change accordingly to avoid mismatches between the image and odors. In this step, ChatGPT accesses relevant information from the short-term storage unit of the previous frame, including people, objects, scenes, and emotional characteristics in the frame, and compares it with the current frame. If no significant differences are found, the odor category for this dimension will not be adjusted. Because ChatGPT's selection of odors has a degree of randomness, this ensures a coherent experience for viewers without experiencing confusing odor information in the same scene.
- Emotion: If the emotional characteristics of this frame change,

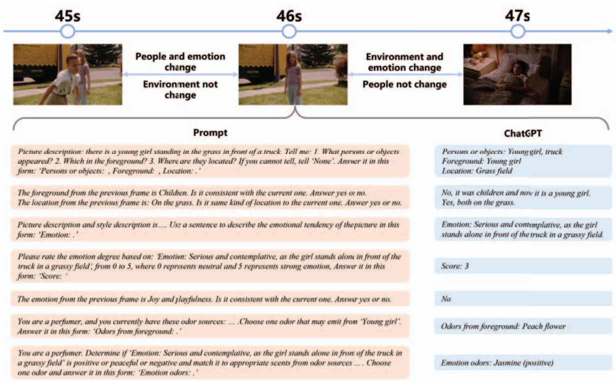

Figure 2: Example of odor sequence generation based on proposed odor matching frame.

ChatGPT needs to recall the description of these emotional characteristics from the ”Global” context, determine whether this emotional tendency is positive, neutral, or negative, and propose suitable odors from the odor library. Simultaneously, a score is assigned to the intensity of this emotional tendency, which determines the concentration of odors associated with the emotion.

- Main Object: If people, objects, or scenes appearing in this frame change, ChatGPT needs to recall this information from ”global,” using its associative capabilities to propose the possible odors that may be emitted from the frame based on the odor library.

- Space: Individuals and objects in the frame are located closer to the audience, while the scene as the background is farther away. By invoking ZoeDepth, the distances of the foreground which is the average distance of smaller distance half, and background which is the average distance of greater distance half are obtained, allowing for adjustments in the concentration differences of different odors.

Proportional odors: The OdorAgent will match up to three different odors for each frame to provide viewers with an immersive experience. Mixing too many odors can make viewers uncomfortable. Among these three odors, one is related to the foreground of the frame, one is related to the background of the frame, and one is related to the emotional context of the frame. The relative concentrations of these three odors are determined based on the distance between the foreground and background and the intensity score of the emotion. In the generated result, the foreground of the frame is typically the core narrative content, and the addition of background odors complements the emotional odors, creating a better atmosphere.

Considering the speed of information processing, we capture frames of a movie segment every 1 second, as shown in Fig. 2. This segment tells the story of a boy running towards his mother due to shyness, leaving the girl alone on the grass. From a viewer’s perspective, we can discern the difference between the 45th and 46th seconds of the scene. We notice the characters in the frame changed, no change in the environment, and the emotional context changed. It is evident that ChatGPT can comprehend the scene and accurately detect changes in the main subject, location, and emotional tone. Ultimately, for the young girl in the 46th-second frame, ChatGPT selected the scent of peach flowers, selected the scent of jasmine to enhance the emotional ambiance, and continued using the scent of grass for the environmental atmosphere from the 45th-second frame.

Figure 3: 88 odor sources used in the experiment.

4 Experiment and Result

While previous research has indicated that incorporating odors enhances user immersion, because large language models is relatively new, there still remains some research questions to address regarding the use of large language models for generalized intelligent olfactory experience design. Therefore, the primary purpose of conducting experiments is to assess whether the OdorAgent framework, which utilizes large language models for odors allocation, can appropriately match odors for different scenarios and systematically evaluate the roles of the three olfactory features within OdorAgent, thereby obtaining valuable user feedback through this exploratory experiment.

4.1 Preparation

We selected the movie ”Flipped” and captured a segment lasting 2 minutes and 40 seconds. The reason for selecting this particular movie segment is due to its inclusion of three distinct settings: outdoor panoramic views, outdoor close-ups view, and indoor scenes. Additionally, this segment narrates the shifting emotional state of the female protagonist, encompassing emotions such as affection, anticipation, disappointment, tranquility, and acceptance. This segment includes the following scenes:

- Scene 1: The young girl’s first encounter with the boy she likes during her childhood, leading to love at first sight.

- Scene 2: The young girl lies on her bed, reminiscing about this experience and feeling regretful about being interrupted by the boy’s mother while recalling her time with the boy.

- Scene 3: As she grows up, the young girl finds it difficult to move on because she never dated the boy she liked. She receives guidance from her father to help her cope with this.

- Scene 4: The girl frequently climbs trees and gazes into the distance, gradually understanding her father’s words and finding solace.

We have acquired 88 different aromatic odor essences (Fig. 3, from the Charm Kaiser Smell Museum). These odor essences encompass various categories such as floral, snacks, fruits and vegetables, beverages, vegetations, scenarios, and daily necessities. We have provided ChatGPT with labels for these 88 aromatic essences, enabling ChatGPT to select suitable odors from this pool to complement the visuals. The reason for selecting these particular fragrance essences is because they fully meet the requirements of this movie segment. Additionally, purchasing these odor essences from the same vendor ensures consistency in the overall olfactory experience.

However, when using the OdorAgent to match odors to video frames, despite having a well-structured framework, unusual situations can still occur. For example, the OdorAgent may select odors not present in the provided odor library, or it may provide different results after each run. To clarify which odors are needed for the user experiment, we ran the OdorAgent multiple times and selected the 16 most frequently occurring odors from the results: Gardenia, jasmine, strawberry ice cream, espresso, wild chrysanthemum, cigar, baby powder, paperback, grass, fresh toothpaste, dust, lavender, bits of wood, ocean, forest, and bamboo. These 16 odors were then used as a new odor library to match odors with video frames.

To investigate the effectiveness of the OdorAgent framework, we designed four comparative experiments:

- A. Blank Control Group.
- B. FO (Foreground Odor), representing the scent associated with the main content displayed in the frame.
- C. FO + BO (Background Odor) + EO (Emotion Odor), introducing background odors, including environmental odors and emotional odors, alongside the main content in the frame. All three odors are played at the same concentration.
- D. FO + BO + EO, where the proportion of the three odors are adjusted based on distance and intensity.

By conducting these four sets of experiments, we aim to assess how odors associated with the main content, odors associated with the background and emotions, and spatial distance adjustments within the scent design framework impact the overall olfactory experience.

In this study, we employed ultrasonic atomization to deliver odors and created a 4x4 atomization array to provide a multimodal viewing experience for VR goggles. We bound the 16 aromatic essences, diluted in a 1:30 ratio, to ultrasonic atomization chips and controlled each chip using a Mega 2560 Pro to play specific odors. This approach offers the advantage of adjusting the atomization intensity by modulating the vibration duty cycle of the atomization chips, thereby generating mixed odors with concentration differences.

In previous research, issues such as the speed of odor transitions and the lingering of odors have been mentioned. To achieve rapid odor transitions and dispersion, during the experiment, we placed a small fan approximately 1 meter away from the odor display device. This fan is used to blow the odors towards users wearing VR goggles and watching movie clips while facing the odor display device, ensuring the timely delivery of odors feedback.

4.2 Participants

A total of 16 participants took part in this experiment, comprising 7 males and 9 females, with ages ranging from 19 to 30 years (M=23.06, SD=3.21). Among them, 5 participants had prior experience in fragrance blending. We recruited these volunteers by distributing promotional posters and obtained their informed consent. Before commencing the experiment, we asked each volunteer to test their olfactory discrimination ability, with the criterion being the ability to accurately distinguish three different odor sources after smelling them for 10 seconds.

4.3 Procedure

4.3.1 Test 1

After confirming that the participants' olfactory discrimination abilities were satisfactory, they were required to wear VR goggles and sit at a table. In front of each participant was placed a prototype of the odor display device with 16 channels that we had prepared. Positioned 1 meter away from the participants, there was a weak fan capable of directing the odors released by the odor display device towards the participants. The airflow from the fan was subtle enough not to cause any noticeable breeze for the participants.

Following the experimental design of the comparative study, participants watched the selected "Flipped" movie clip four times. The four modes included: A. Blank Control Group. B. FO. C. Equal proportion of FO+BO+EO. D. Adjusted proportion of FO+BO+EO. These four modes were presented in a randomized order to eliminate any potential sequence effects. After each viewing, participants

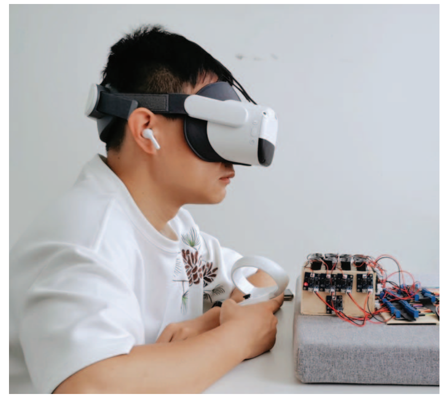

Figure 4: Participant is watching segment accompanied by odors.

filled out "The Film IEQ" questionnaire [44, 32, 40], which assessed their viewing immersion, and they also sniffed coffee beans to restore their sense of smell. After completing the four repetitions of the movie clip, participants were asked to rank their experiences for the four viewings.

4.3.2 Test 2

The OdorAgent can only match odors to videos based on odor descriptors. Therefore, after participants rated the match between each odor and its odor label, they directly viewed images captured by the OdorAgent along with the corresponding odor text labels. From the perspective of the text labels, participants assessed the odors selected by the OdorAgent and determined how well they matched the images extracted from the video for each second. Participants were required to rate the match between odors and visuals for the three different odor matching methods (B, C, D).During this phase, to reduce the workload for all participants, the time interval for displaying images to participants was set at 5 seconds.

Finally, we conducted semi-structured interviews with the participants. The questions included: 1. Which segments do you think had good matches between odors and visuals? 2. Which segments do you think had poor matches between odors and visuals? 3. How did your experience compare between segments C and D, where the only difference was in odors proportion?

4.4 Result

4.4.1 Test 1

In this experimental study, a total of 16 participants were invited to engage in the actual viewing of film clips incorporating four distinct olfactory blending approaches. The evaluation scale for this experiment was "The film IEQ" questionnaire [44, 32, 40]. The data underwent a Lilliefors test, confirming its adherence to a normal distribution ($p>0.05$). Simultaneously, a test for homogeneity of variances was conducted ($p=0.50>0.05$), indicating that the data adhered to a normal distribution with equal variances. Therefore, a one-way analysis of variance (ANOVA) was performed on the data.

According to the results of the questionnaire, it was observed that, overall, the introduction of olfactory stimuli had a noticeable enhancement on the immersive experience of users utilizing VR devices for film viewing. However, the differences between various olfactory stimulus schemes did not appear significant. Specifically, Group D, which involved a carefully adjusted concentration ratio of FO+BO+EO, received the highest rank(Fig. 6) in the tests (M=4.984, SD=0.763). This was nearly identical to Group C, which employed a scheme featuring equal concentrations of FO+BO+EO (M=4.987, SD=0.526). Group B, which featured only background olfactory stimuli, scored slightly lower than the aforementioned two groups (M=4.844, SD=0.691). The control group with no olfactory stimuli

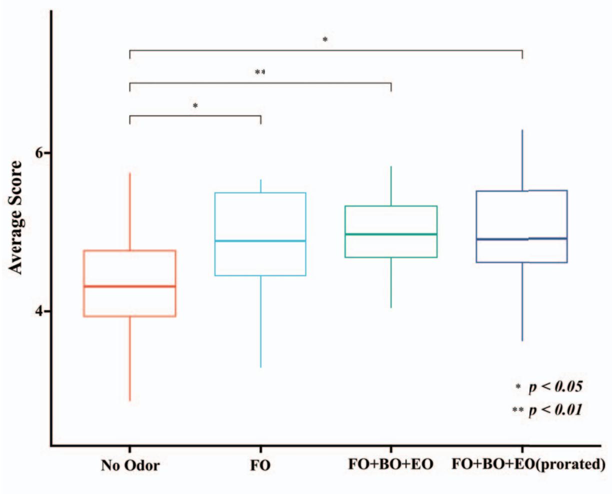

Figure 5: Average score of test 1.

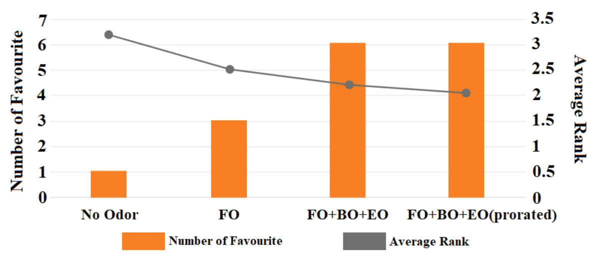

Figure 6: Sort the different odor sequences according to the degree of preference, the number of categories that are ranked first, as well as the average ranking for each category.

(M=4.289, SD=0.747) displayed significantly lower ratings compared to the three olfactory stimulus groups.

Furthermore, the results of the significance test conducted on the four groups (F=3.51, $p<0.05$) indicated clear significance. Subsequent analysis of significance between Groups A&B (F=4.46, $p<0.05$), Groups A&C (F=8.76, $p<0.01$), and Groups A&D (F=6.36, $p<0.05$) demonstrated that Groups B, C, and D significantly outperformed Group A without olfactory stimuli. Conversely, the differences among Groups B, C, and D were not statistically significant (F=6.46, $p>0.05$).

Based on the comprehensive analysis of the statistical results above, we can draw the following conclusion: Olfactory stimulation enhances immersion in film viewing for individuals; however, people do not exhibit sensitivity to specific olfactory stimuli. From a variance perspective, when comparing Group C and Group D, participants' evaluations in Group D displayed relatively greater variance, indicating more significant divergence in their responses.

At the end of the four-round comparison experiment, participants had been asked to rank the four viewing experiences. We counted the number that were ranked first in satisfaction and calculated the average rank for each experience. The results(Fig. 6) showed that users were most satisfied with Groups C and D, and second most satisfied with Group B. This is generally consistent with the results obtained through the standard questionnaire.

4.4.2 Test 2

In this experimental project, we evaluated the preferences of 16 participants for four different odor blending schemes, with results obtained from a 7-point scale. Similarly, after conducting a Lilliefors

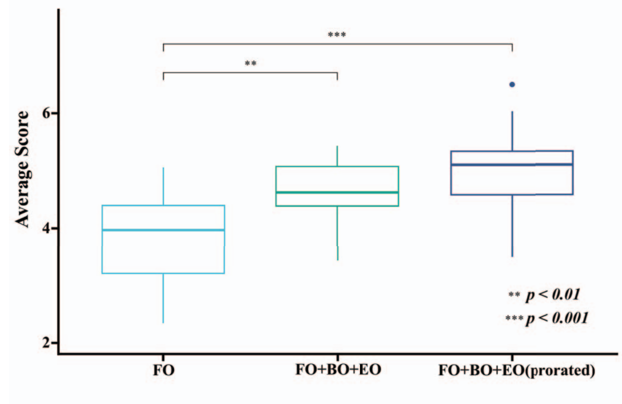

Figure 7: Average score of test 2.

test ($p>0.05$) and a test for homogeneity of variances ($p=0.45>0.05$), we proceeded with a one-way analysis of variance (ANOVA). The result is shown in Fig. 7.

Ultimately, the odor blending scheme in Group D received the highest scores (M=5, SD=0.764), with Group C's scheme (M=4.625, SD=0.549) slightly below, consistent with the results from the olfactory stimulation experiment. Group B alone (M=3.832, SD=0.766) ranked third. Significance testing conducted on these results (F=12.25, $p<0.05$) indicates that this outcome holds statistically significant meaning. Further pairwise significance testing reveals that the differences between Groups C&D are not significant (F=2.38, $p>0.05$). Meanwhile, Groups B&C (F=10.61, $p<0.01$), Groups B&D (F=17.46, $p<0.001$).

The results indicate that users tend to favor odor blending schemes that incorporate more visual information, but the evaluation improvement from altering the olfactory concentration ratios is not significant.

4.4.3 Analysis of Interview

The majority of participants expressed their fondness for this experience, considering the scented version to be a superior experience compared to the version without odors. Most participants specifically mentioned that the olfactory experience in the first scene was exceptional. Whether it was the timing of the odor release or the ambiance it created, it significantly captured their attention and enhanced immersion.

For the second scene, where the young girl lay on the bed reminiscing about her encounter with the young boy, four participants also found this scene satisfying. They felt it portrayed the girl's innocence and emotions effectively.

In the scene where the girl chats with her father, seven participants found the odor in this scene less satisfactory. Their feedback primarily fell into two categories: some thought the addition of the coffee scent did not match the visuals, while others believed the overall odor style leaned towards sweetness.

Regarding the final scene of climbing the tree, there were mixed opinions about the odor evaluation. Seven participants found the odors in this scene refreshing, while four participants believed the odors did not align well with the visuals.

Overall, participants showed a relatively high level of acceptance for the odor content. The lower ratings for the scene where the girl chats with her father may have two main reasons. Firstly, ChatGPT associated the coffee scent with the father, and although there might be a connection, it didn't quite match the scene as they were in the wilderness rather than an office setting. Secondly, most of the fragrances we purchased tended to have sweet notes, which did not align well with the relatively serious conversation in that scene. Similar issues were observed in the tree-climbing scene, where

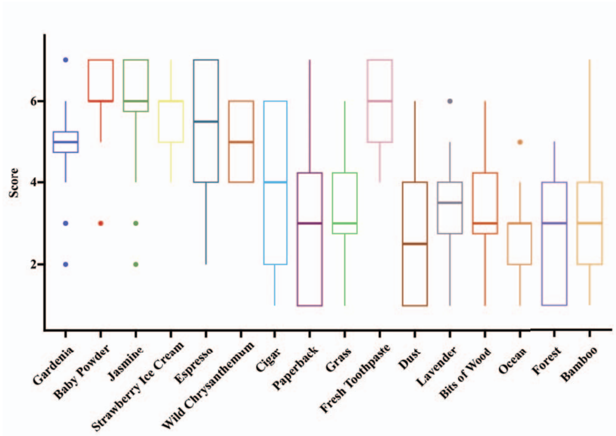

Figure 8: Score of matching between Odors and Labels.

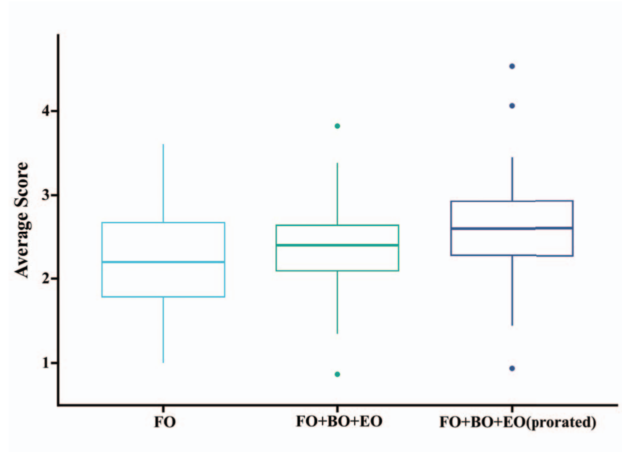

Figure 9: Adjusted average score of Test 2.

the intended fresh and calming scents were mixed with a cloying sweetness, making it feel out of place for some participants.

Participants generally perceived the odors in Group C as milder and the odors in Group D as more intense. This observation could be attributed to the adjustment of odor ratios in Group D, which made specific odors more pronounced in that group.

5 Supplementary Experiment and Expert Interviews

5.1 Supplementary Experiment

As the aromatic essences we purchased may not perfectly align with people's cognitive expectations regarding their specified scent categories, this may cause the difference results between Test 1 and Test 2.

5.1.1 Procedure

To eliminate the influence of the irrelevant variable of "match between scent name and actual perception," each participant was required to, after completing all four repetitions of the movie clips, smell each of the aromatic essences we used. Participants were given 10 seconds for each essence to experience and contemplate it (smell + think). Subsequently, participants were asked to rate the accuracy of the descriptions provided for each aromatic essence on a scale from 1 to 7, where 7 indicates a complete match between perception and description, and 1 indicates no match at all. There was a 30-second interval between each rating to allow participants to reset their olfactory senses.

5.1.2 Result

In the film-viewing experiment, the overall effectiveness of olfactory stimulation can be influenced by the quality of the odor used. Based on the data analysis presented earlier, there is a difference between the ratings given for actually watching scented films and the ratings based solely on the odor labels. To investigate whether this difference is attributable to the quality of the odors, we conducted further analysis by having participants rate the matching degree between the 16 odors used and their respective labels. This was done using a 7-point scale, where 7 indicates a perfect match, and 1 signifies no match at all. The results(Fig. 8) reveal that not all odors can deliver the olfactory experience claimed by the manufacturers, and this can indeed impact the user experience in the final practical tests.

The matching scores between the visual content and odor labels indicate that participants rated Groups C and D significantly higher than Group B. However, in the actual viewing of scented films, the differences between Groups B, C, and D were not significant. To explain this disparity, we attempted to adjust the results of Test 2 based on participants' matching score between odors and labels. The goal was to eliminate the impact of differences between the odors and the information claimed by the manufacturers. Test 2 represents an ideal scenario where all odors should match people's memory of odor experiences. Therefore, the core of this adjustment is to use the matching score of odor labels as coefficients multiplied by the ratings of Test 2 to reflect the real situation.

Score corrections were made by multiplying each participant's Test 2 score by that participant's own scoring of the odor and label match, because each person's experience of an odor is subjective. For example, some people thought that the lavender odor we supplied matches well to the odor he perceived to be lavender, but others thought that the odor reminded them of laundry detergent, which affected their judgment.

After conducting the Lilliefors test ($p>0.05$) and the test for homogeneity of variances ($p=0.60>0.05$) on all The result is shown in Fig. 9. The highest average score is still observed in Group D ($M=2.651$, $SD=0.867$), followed by Group C ($M=2.378$, $SD=0.726$), and Group B ($M=2.24$, $SD=0.67$) with the lowest score. This ranking aligns with the results from Test 2. However, unlike Test 2, the differences between the three groups are not significant in this case ($F=1.14$, $p>0.05$), which is consistent with the non-significant differences observed between Groups B, C, and D in Test 1.

Hence, we can conclude that the difference between the results of the odor rating experiment and the scented film viewing experiment is attributed to the matching quality of the odor elements themselves. During the actual viewing of scented films, the significant impact on the results is due to the disparities between the actual odors and their claimed attributes in the labels.

5.2 Expert Interviews

In order to determine the validity of the scent sequences generated by OdorAgent, we invited three experts in the field of scent to have a discussion, with their areas of expertise being olfactory human-computer interaction design, perfume blending, and scent healing. The purpose of the discussion was to obtain the experts' opinions on generating odour sequences for the videos, and to ask the experts to also match odours to the clips used in the experiments for visual comparison.

The experts agreed that scent pairing for video to produce a multimodal experience is an interesting topic and there are many points worth investigating in terms of experience. The discussion led to a number of valuable conclusions: 1. The insertion of odors needs to depend on whether the focus of the content of the video material is the visual experience or the plot; 2. There are odors when there is a scene, and the environmental information brought by the scene, the sense of immersion, followed by the odor in the foreground, and then the mood and atmosphere; 3. Attention needs to be paid to the utilisation of the line to understand the content of the video; 4. Playing the odor in the foreground or in the background has less to do with the screen distance has little to do with it, and the plot has more to do with it; 5. the taste can not be too real, too real taste may have a bad experience, such as stench; 6. need to pay attention to the colour mixing, filters, and the impact on the emotion of the picture; 6. need to pay attention to the use of sound information, such as in the use of odors on behalf of the person, when his voice appeared, then you can add the odors associated with him. The experts also predicted that in the future it might be possible to add odour information directly when filming, but for the time being it can be used in such a way as to be compatible with past films OR videos. From the results of the discussion, the framework proposed by OdorAgent was accepted by the experts, but there is still room for improvement in the use of some detailed information.

A major difference between the expert-matched odour sequences and the results generated by OdorAgent is that the expert-matched odours are not as varied and do not change as frequently. The experts selected scents based more on an overall understanding of the scene than the framework used by OdorAgent which distinguishes between foreground, background and mood, sometimes focusing on the smells of the objects in the scene, and sometimes focusing on the overall feel of the scene. For example, the results of the experts' scent matching for Scene 3 show that they focused more on the ambient scent in this scene because they thought that the ambient scent would be sufficient for this scene because it did not have a very strong emotion, whereas OdorAgent still associates a scent based on the style of the scene.

6 Discussion

6.1 Feasibility

Due to its unique nature, there hasn't been a satisfactory digital solution for odors. However, owing to the close connection between the human olfactory sense and other senses, it suggests the possibility of utilizing large language models for odor generation and matching. We designed rigorous user experiments to verify whether using a large model to generate odor compositions matching film clips could meet user needs. The results demonstrated that odor compositions generated using the OdorAgent significantly enhanced the user's viewing experience. Among the different generation methods using the large model, those that incorporated the main subjects of the scene, descriptions of the scene's style, and spatial information of the scene yielded significantly higher user satisfaction than relying solely on the main subjects of the scene for odor composition. However, whether or not spatial information was used did not have a particularly significant impact on the user experience.

We surmise that this is because human olfactory perception is not as precise as visual or auditory senses; it is more of a vague sensation. Therefore, even though we used the OdorAgent to identify depth information in images and adjust odor concentration accordingly, some users still couldn't perceive these differences. Nevertheless, optimistically, some odor-sensitive users were able to detect these distinctions, encouraging us to work towards more precise inference of odor features for output.

The results of the user experiments also demonstrate that OdorAgent possesses a degree of scene versatility, allowing inexperienced users to easily and conveniently design olfactory experiences for video scenes. For VR device, which emphasises user immersion, OdorAgent adds the possibility of odour modalities for more applications in VR in the future, such as VR games or VR theater. However, the actual effectiveness is notably influenced by the purchased fragrances. Expert users can utilize OdorAgent to assist them in designing satisfying olfactory experiences for video scenes, leveraging their professional expertise to create well-matched odors. The work in this paper focuses more on verifying whether the design framework in OdorAgent has personalized potential. OdorAgent holds the potential for personalization, and based on our framework, it can also facilitate personalized prompts, thus yielding odor sequences that align more closely with individual preferences and habits.

6.2 Limitations

We utilized the CLIP Interrogator model, Zoedepth model, and ChatGPT model for odor modulation and pairing, but these three models were not originally designed for this specific task. In terms of image recognition results, there may be some minor errors in recognizing certain images, but they generally do not affect the primary content conveyed by the visuals. Based on user interviews, ChatGPT tends to match odors to as many images as possible, which could lead to olfactory fatigue for users during extended viewing. Additionally, questionnaire results indicate that some users prefer scenes with no odors.

ChatGPT is skilled at using associative and imaginative abilities to link odors with objects appearing in the images. This can lead to two possible outcomes: if a user's memories and associations align with or are similar to ChatGPT's associations, their experience will be favorable. However, if a user's understanding and memories differ from ChatGPT's results, they may perceive unrelated odors being played. This underscores the need for us to prompt ChatGPT as effectively as possible during the process to ensure it provides the most fitting associative results. There is some randomness in the results generated by ChatGPT, but it is within reasonable limits. For example, the floral fragrances used in the experiment appear to be somewhat generic to ChatGPT, and could be chosen by anyone in many scenarios.

We have only conducted one experiment of an exploratory nature, and the problem of generating odour sequences for video does leave much to be explored, with future experiments incorporating more objective parameters.

6.3 Future work

Contextual understanding: Strengthening the model's comprehension of the film's context and making odor deliver decisions based on plot developments can lead to more synchronized odor deliver with visuals over time.

Multi-modal large model: The current implementation is a distributed multi-modal system, which requires two steps, image-text model and language large model, to achieve multi-modal analysis. This results in delay and more computational power consumption. Future work requires further development of native multi-modal models that integrate images, texts, and smells.

Speech-based understanding through dialogue: Utilizing speech recognition and large language models to understand movie plot developments based on dialogue text can enhance the accuracy of recognition and inference.

Deliver rules: Current deliver rules aim to match odors with visuals as much as possible, which can lead to olfactory fatigue for users and hinder their ability to watch odorized films for extended periods. This necessitates the development of more nuanced deliver rules that address olfactory fatigue while enhancing the overall user experience.

Hardware design for odor deliver: Exploring various odor deliver technologies and investigating factors influencing user experiences can lead to advancements in the design of odor deliver devices.

7 Conclusion

Despite the vision-driven creation, we need to address technical challenges in digital scent technologies in HCI. It must be acknowledged that odors cannot be synthesized or algorithmically generated in the same way as visuals up to now. This is one of the reasons why odor display technology has not been widely adopted—the inherent limitations. We have found that large language models can significantly address one of these challenges by enabling us to find ways to approximate movie odors within the limited range of odor categories intelligently and without manual intervention.

We have introduced an OdorAgent capable of generating odor compositions for films, leveraging three models: CLIP Interrogator, Zoedepth, and ChatGPT. We divided the information extracted by the large models from the visuals into three parts: the main subjects in the scene, emotional context, and spatial location. ChatGPT then analyzed this information and selected odors from our prepared set to match the visuals. We recruited 16 volunteers to test the effectiveness of this approach by having them watch film clips with odor deliver. We also discussed the differences between various generation methods. The experimental results strongly validate the OdorAgent's ability, which rivals that of human perfumers and significantly enhances the user's film-viewing experience. After addressing the issue of odor discrepancies with labels, the results indicate that having no odors and relying solely on the main subjects in the scene for odor generation significantly resulted in lower scores compared to the other two generation methods.

The use of the OdorAgent enables automatic generation of odor compositions for film visuals, offering the potential for widespread applications of odorized films. Moreover, the OdorAgent possesses flexible generation rules, allowing it to adapt odors to the main subjects and conveyed emotions in visuals through associations and imagination. This reduces the requirements for the variety of odor sources and can enrich the user experience even with a limited set of odors on odor deliver devices.

In this paper, we employ emerging AI technologies to address the challenges of pairing odors with images and videos (for instance, movies), which traditionally require a high level of expertise. This approach not only offers assistance but also guidance for future research in this domain.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 62172252) and the Guoqiang Foundation (Grant No. 2020GQG0004). We thank all the participants who joined in our user study and expert interviews.

References

[1] S. Alraddadi, F. Alqurashi, G. Tsaramirsis, A. Al Luhaybi, and S. M. Buhari. Aroma release of olfactory displays based on audio-visual content. *Applied Sciences*, 9(22):4866, 2019. 2

[2] J. Amores and P. Maes. Essence: Olfactory interfaces for unconscious influence of mood and cognitive performance. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pp. 28–34, 2017. 2

[3] A. Bahremand, M. Manetta, J. Lai, B. Lahey, C. Spackman, B. H. Smith, R. C. Gerkin, and R. LiKamWa. The smell engine: A system for artificial odor synthesis in virtual environments. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 241–249. IEEE, 2022. 2

[4] O. Baus and S. Bouchard. Exposure to an unpleasant odour increases the sense of presence in virtual reality. *Virtual Reality*, 21:59–74, 2017. 1

[5] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 2

[6] J. Brooks and P. Lopes. Smell & paste: Low-fidelity prototyping for olfactory experiences. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2023. 1, 2

[7] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2

[8] J. P. Cater. Smell/taste: odors in reality. In *Proceedings of IEEE International Conference on Systems, Man and Cybernetics*, vol. 2, pp. 1781–vol. IEEE, 1994. 2

[9] J. A. Fernandez. *Olfactory interfaces: toward implicit human-computer interaction across the consciousness continuum*. PhD thesis, Massachusetts Institute of Technology, School of Architecture and Planning . . . , 2020. 2

[10] G. Ghinea and O. A. Ademoye. Olfaction-enhanced multimedia: perspectives and challenges. *Multimedia Tools and Applications*, 55:601–626, 2011. 2

[11] A. Gilbert. What the nose knows: The science of scent in everyday life. *(No Title)*, 2008. 2

[12] D. Gosain and M. Sajwan. Aroma tells a thousand pictures: digital scent technology a new chapter in it industry. *Int. J. Curr. Eng. Tech*, 4:2804–2812, 2014. 2

[13] M. L. Heilig. Sensorama simulator. *US PAT. 3,050,870*, 1962. 2

[14] R. S. Herz and T. Engen. Odor memory: Review and analysis. *Psychonomic Bulletin & Review*, 3(3):300–313, 1996. 1

[15] A. K. Holloman and C. S. Crawford. Defining scents: A systematic literature review of olfactory-based computing systems. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(1):1–22, 2022. 2

[16] W. IJsselsteijn. Presence in the past: what can we learn from media history? In *Being there: concepts, effects and measurements of user presence in synthetic environments*, pp. 17–40. Ios Press, 2003. 2

[17] L. Jones, C. A. Bowers, D. Washburn, A. Cortes, and R. V. Satya. The effect of olfaction on immersion into virtual environments. *Human performance, situation awareness and automation: Issues and considerations for the 21st century*, pp. 282–285, 2004. 1

[18] S. Kim, J. Park, J. Bang, and H. Lee. Seeing is smelling: Localizing odor-related objects in images. In *Proceedings of the 9th Augmented Human International Conference*, pp. 1–9, 2018. 2

[19] P. Kortum. *HCI beyond the GUI: Design for haptic, speech, olfactory, and other nontraditional interfaces*. Elsevier, 2008. 2

[20] Y. Lin, K. Zheng, L. Liu, Y. Chen, J. Guo, S. Li, C. Yao, and F. Ying. Olfacenhancer: A vision-based scented necklace for cross-modal perception and olfaction augmentation. *Design and Semantics of Form and Movement*, p. 237. 2

[21] E. Maggioni, R. Cobden, D. Dmitrenko, K. Hornbæk, and M. Obrist. Smell space: mapping out the olfactory design space for novel interactions. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 27(5):1–26, 2020. 2

[22] D. Martin, S. Malpica, D. Gutierrez, B. Masia, and A. Serrano. Multimodality in vr: A survey. *ACM Computing Surveys (CSUR)*, 54(10s):1–36, 2022. 2

[23] M. Melo, H. Coelho, G. Gonçalves, N. Losada, F. Jorge, M. S. Teixeira, and M. Bessa. Immersive multisensory virtual reality technologies for virtual tourism: A study of the user's sense of presence, satisfaction, emotions, and attitudes. *Multimedia Systems*, 28(3):1027–1037, 2022. 2

[24] M. Melo, G. Gonçalves, P. Monteiro, H. Coelho, J. Vasconcelos-Raposo, and M. Bessa. Do multisensory stimuli benefit the virtual reality experience? a systematic review. *IEEE Transactions on Visualization and Computer Graphics*, 28(2):1428–1442, 2020. 2

[25] N. Murray, O. A. Ademoye, G. Ghinea, and G.-M. Muntean. A tutorial for olfaction-based multisensorial media application design and evaluation. *ACM Computing Surveys (CSUR)*, 50(5):1–30, 2017. 2

[26] N. Murray, B. Lee, Y. Qiao, and G.-M. Muntean. Olfaction-enhanced multimedia: A survey of application domains, displays, and research challenges. *ACM Computing Surveys (CSUR)*, 48(4):56, 2016. 1, 2

[27] T. Nakamoto. *Human olfactory displays and interfaces: Odor sensing and presentation*. IGI Global, 2013. 2

[28] T. Nakamoto, T. Hirasawa, and Y. Hanyu. Virtual environment with smell using wearable olfactory display and computational fluid dynamics simulation. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 713–720. IEEE, 2020. 2

[29] T. Nakamoto, S. Otaguro, M. Kinoshita, M. Nagahama, K. Ohinishi, and T. Ishida. Cooking up an interactive olfactory game display. *IEEE Computer Graphics and Applications*, 28(1):75–78, 2008. 1

[30] T. Narumi, S. Nishizaka, T. Kajinami, T. Tanikawa, and M. Hirose.

Augmented reality flavors: gustatory display based on edible marker and cross-modal interaction. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 93–102. ACM, 2011. 1
[31] S. Niedenthal, W. Fredborg, P. Lundén, M. Ehrndal, and J. K. Olofsson. A graspable olfactory display for virtual reality. *International journal of human-computer studies*, 169:102928, 2023. 1, 2
[32] K. Nottingham, P. Ammanabrolu, A. Suhr, Y. Choi, H. Hajishirzi, S. Singh, and R. Fox. Do embodied agents dream of pixelated sheep?: Embodied decision making using language guided world modelling. *arXiv preprint arXiv:2301.12050*, 2023. 2, 5
[33] M. Obrist, C. Velasco, C. Vi, N. Ranasinghe, A. Israr, A. Cheok, C. Spence, and P. Gopalakrishnakone. Sensing the future of hci: Touch, taste, and smell user interfaces. *interactions*, 23(5):40–49, 2016. 1
[34] D. Panagiotakopoulos, G. Marentakis, R. Metzitakos, I. Deliyannis, and F. Dedes. Digital scent technology: Toward the internet of senses and the metaverse. *IT Professional*, 24(3):52–59, 2022. 2
[35] N. K. Parashar and N. Kumar. A concept of digital scent/smell technology: an underrated technology. *Int J Recent Innov Trends Comput Commun*, 5(11):254–256, 2017. 2
[36] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*, 2023. 2
[37] M. Paterson. Digital scratch and virtual sniff: Simulating scents. *The smell culture reader*, pp. 358–367, 2006. 2
[38] N. Ranasinghe, P. Jain, N. Thi Ngoc Tram, K. C. R. Koh, D. Tolley, S. Karwita, L. Lien-Ya, Y. Liangkun, K. Shamaiah, C. Eason Wai Tung, et al. Season traveller: Multisensory narration for enhancing the virtual reality experience. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pp. 1–13, 2018. 1, 2
[39] M. Rheiner. Birdly an attempt to fly. In *ACM SIGGRAPH 2014 Emerging Technologies*, pp. 1–1. 2014. 2
[40] J. M. Rigby, D. P. Brumby, S. J. Gould, and A. L. Cox. Development of a questionnaire to measure immersion in video media: The film ieq. In *Proceedings of the 2019 ACM International Conference on Interactive Experiences for TV and Online Video*, pp. 35–46, 2019. 5
[41] S. A. Seah, D. Martinez Plasencia, P. D. Bennett, A. Karnik, V. S. Otrocol, J. Knibbe, A. Cockburn, and S. Subramanian. Sensabubble: a chrono-sensory mid-air display of sight and smell. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2863–2872, 2014. 2
[42] S. Sugimoto, D. Noguchi, Y. Bannnai, and K. Okada. Ink jet olfactory display enabling instantaneous switches of scents. In *Proceedings of the 18th ACM international conference on Multimedia*, pp. 301–310. ACM, 2010. 2
[43] G. Tsaramirsis, M. Papoutsidakis, M. Derbali, F. Q. Khan, and F. Michailidis. Towards smart gaming olfactory displays. *Sensors*, 20(4):1002, 2020. 2
[44] Z. Wang, S. Cai, A. Liu, X. Ma, and Y. Liang. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. *arXiv preprint arXiv:2302.01560*, 2023. 2, 5
[45] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. 2
[46] J. Willander and M. Larsson. Smell your way back to childhood: Autobiographical odor memory. *Psychonomic bulletin & review*, 13(2):240–244, 2006. 1
[47] Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou, et al. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*, 2023. 2
[48] Z. Yang, L. Li, J. Wang, K. Lin, E. Azarnasab, F. Ahmed, Z. Liu, C. Liu, M. Zeng, and L. Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023. 2
[49] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022. 2
[50] H. Zhang, C. Wu, J. Xie, C. Kim, and J. M. Carroll. Qualigpt: Gpt as an easy-to-use tool for qualitative coding, 2023. 2