Wish3D, Mixed Reality Object Generation from Speech

Oikonomou Orestis[1], Raphael Winkler[2], Dochev Grigor Dobromirov[1], Saxena Rupal[1], Philipp Lindenberger[23]

[1]Department of Informatics, University of Zurich, CH
[2]Department of Informatics, ETH Zurich, CH
[3]Supervisor

Abstract

This report presents 'Wish3D', a Mixed Reality (MR) application designed for the Magic Leap 2 Headset. Wish3D utilises advanced speech recognition and 3D generation technologies to enable users to create and interact with 3D objects through voice commands in a Mixed Reality environment. Users can engage with the produced objects within the environment, having the ability to rotate, relocate, adjust scale, and throw objects in MR. Notably, objects are capable of interacting with one another through collisions and aligns with the principles of physics governed by gravity.

Wish3D demonstrates a significant advancement towards intuitive and interactive Mixed Reality experiences, despite some limitations in its ability to generate high quality objects. The application's real-time performance is evident by an average object generation time of 10 seconds, highlighting the potential of generative models being integrated in Mixed Reality environments.

1. Introduction

In the world of interactive technologies, Mixed Reality (MR) stands alone enabling immersive experiences in a blend of the real and virtual world. This report presents *Wish3D*, a MR application for Magic Leap 2 device. *Wish3D* transforms input voice commands into interactive 3D objects in MR. Wish3D leverages spatial mapping of user's surrounding which enables interactions of 3D objects with real world. Users can interact with 3D objects in Wish3D playground. They can scale, rotate, move, pick and place, throw, collide objects in Mixed Reality. Since gravity is applied to all generated 3D objects, they behave as per principles of physics governed by gravity.

Wish3D pipeline integrates speech-to-text conversion with 3D model generation. OpenAI's Whisper model [6] is utilised for translating spoken languages into text. This text is then used as input to the Shap-E model [2], which generates 3D meshes corresponding to the transcribed text. These technologies are integrated with the Magic Leap 2 platform, exemplifying the synergy between advanced machine learning models and 3D rendering capabilities in the field of MR.

The *Wish3D* application was evaluated in a user study involving 20 participants, comprising 4 experienced users and 16 new users. A formative study was conducted with 6 users, and any issues identified during this study were addressed before the presentation of the *Wish3D* application on demo day. To evaluate *Wish3D*, we ask the research question: *"How accurately and effectively does the 'Wish3D' mixed reality application convert voice commands into corresponding 3D models? How does the completion time of user interactions with 3D objects in the 'Wish3D' reflect on the system's usability?"*

2. Background Research

The backbone of the *Wish3D* application is a complex pipeline that integrates speech-to-text conversion with 3D model generation. Voice commands allow more natural experience and with the advancements in the field of speech recognition over the last few years, many MR applications have made use of it [1]. OpenAI's Whisper model [6] is utilised for translating speech into text. This text was then used as input of Shap-E model [2], which generates 3D meshes corresponding to the transcribed text. 3D Object Generation Models such as One2345 [4], Zero123++ [7] with Gaussian Splatting [3], and SDXL [5] were also experimented but not integrated to *Wish3D* because of high inference time as shown in Table 3.

3. Methodology

To enable near real-time inference, we offload all compute-heavy operations to a server, this way we can leverage the more powerful hardware to decrease compute times drasti-

cally. Figure 1 shows the overview of Methodology used to develop *Wish3D* application.

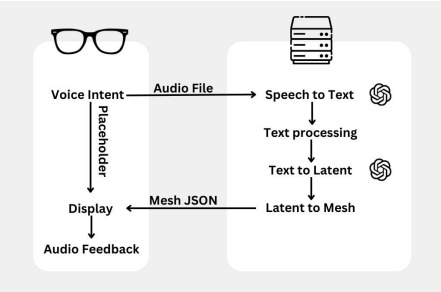


Figure 1. Overview of Methodology

3.1. API

Our backend is built using FastApi, a high performance python web framework. It receives a wav file containing the voice command, and returns one or more 3d objects as json.

3.2. Speech-to-Text

We tested multiple publicly available speech to text models to find a fast and reliable solution to be integrated into our back end. Our final pipeline uses the whisper model [6] by OpenAI due to it's high accuracy. In contrast to other models like S2T [8] which offer better inference times, but vary heavily in their accuracy from user to user. In the pipeline, following the generation of text using the Whisper model, we employ regular expressions to post-process the text.

3.3. Text-to-3D

The field of Text-to-3D, and image-to-3D has made continuous progress over the past years, especially in high quality novel-view synthesis from a single view through diffusion. Unfortunately, it was found that these improvements do not yet translate directly to generating high quality 3D objects.

Many methods based on SDXL [5] claim to be multi-view consistent. Unfortunately, the geometry obtained from this method contained a lot of artefacts. It was noticed that the generated multi-view images are close to consistent, but still contain too many inconsistencies to get reliably good 3D objects.

A more recent approach to generate high fidelity 3D objects in reasonable time is Gaussian splatting [3]. Unfortunately, the fastest publicly available implementation found is still prohibitively slow at around 90 seconds inference time, but the results looked promising in terms of quality. One2345 [4] typically generates slightly better objects, but not enough to warrant the much higher inference time of around 45 seconds.

Therefore, shap-E [2] was finally integrated in *Wish3D*. It is a fast diffusion model that generates a 3D object directly from text, and runs comparatively faster (typically around 4 seconds) due to it's small size. The reduced model size comes with significant compromise in terms of quality of the generated 3D objects.

4. App Instances

This section presents key instances of *Wish3D* application, showcasing its capabilities and user interaction scenarios. The user journey in *Wish3D* begins with the main menu, as shown in Figure 2. This allows users to start or quit the application.

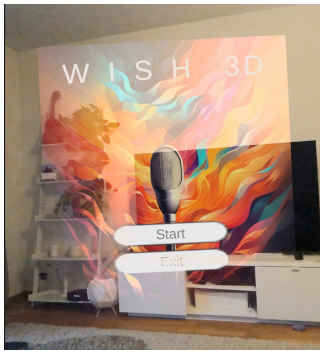


Figure 2. Main menu interface of *Wish3D*

4.1. Onboarding Experience

Once the user selects the *Start* option, an Onboarding session begins. The initial step involves room mapping to ensure accurate interaction with 3D objects in the mixed reality environment, as shown in Figure 3.

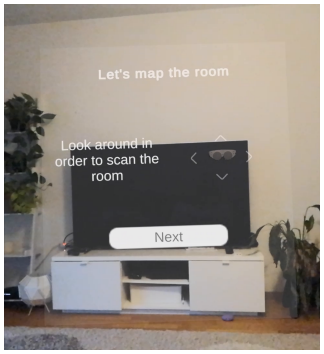
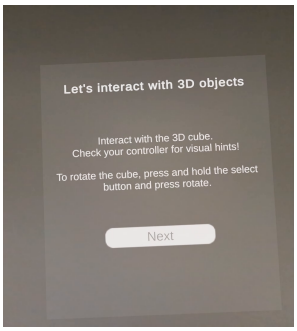
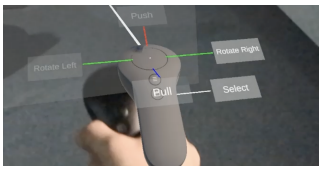


Figure 3. Map room in *Wish3D*

Once the room is mapped, users can learn to interact with a 3D cube in *Wish3D*. They familiarize themselves with controllers. This interaction allow users to understand fundamental interactions within the app. Figures 4a and 4b highlight these key initial interactions.



(a) Learn to Interact with objects



(b) Familiarize with the controller

Figure 4. Guided tour to Interact with 3D Object in *Wish3D*

Users then learn to interact with the application using voice commands. Figure 5 illustrates UI, where users are guided to trigger the app to generate 3D objects through specific voice commands. This step includes exploring various use cases, such as the creation of multiple objects of different types.

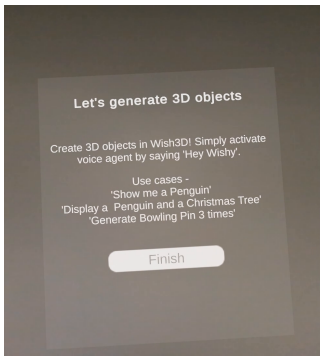


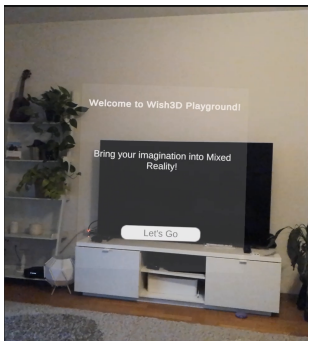
Figure 5. Learn to generate 3D objects from voice command

4.2. Wish3D Playground

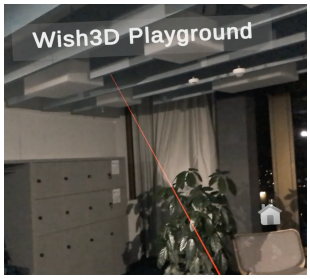
After completing the Onboarding session, users are welcomed into the immersive *Wish3D* playground, as depicted in Figure 6a.

A notable feature of the playground is the inclusion of a new home icon, illustrated in Figure 6b. This icon is placed to inform users that they can return to the main menu at any point, enhancing the overall user navigation experience. It facilitates seamless transitions between the tutorial section and the various functionalities available within the main application.

Within the playground, users have the opportunity to interact with two pre-generated objects, "a chicken with red hat" and "a penguin with bow-tie" as shown in Figures 7a and 7b. These objects are created using One2345 model, showcasing the app's capability to incorporate high-quality 3D objects. This integration exemplifies the potential of the



(a) Users welcomed to the *Wish3D* playground



(b) *Wish3D* playground with Home icon on the left for navigation

Figure 6. *Wish3D* Playground

(a) Chicken with red hat

(b) Penguin with bow-tie

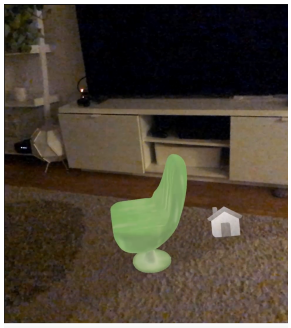
Figure 7. High quality pre-generated objects in *Wish3D* playground

Wish3D app, to continually include diverse new models.

4.3. Object Generation and Interaction

In *Wish3D* playground, users have the freedom to generate and interact with various imaginative objects. As users wait for their creations to materialize, a pop-up cube with a question mark appears on the screen, serving as a loading indicator as shown in Figure 8a, enhances user experience by providing visual feedback during the 3D object generation process. Figure 8b showcases an example of creative object generation, where "A chair that looks like an avocado" was generated by *Wish3D*. Additionally, the application enables users to simulate games like bowling, as shown in Figure 9. This feature not only showcases the app's versatility in creating and interacting with diverse 3D objects but also highlights the sophistication of its physics engine. Through this game play, users can explore the nuances of physics, including collisions and gravity, thereby enriching their overall interactive experience.

(a) Loading indicator



(b) Avocado chair

Figure 8. Features in *Wish3D*: (a) Loading indicator - a pop-up cube with a question mark, and (b) Creative object generation - avocado chair.

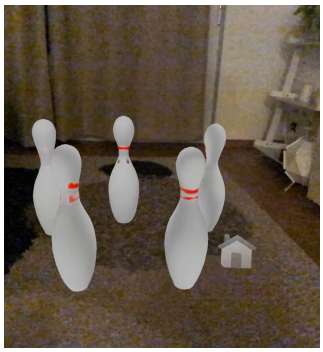


Figure 9. Bowling game setup in *Wish3D*.

5. User Studies

To assess *Wish3D* application, user studies were carried out on 4 experienced users (developers of *Wish3D*) and 16 new users. A Formative User Study was performed on 6 out of 16 new users before the final demo day. The issues appeared in Formative User Study were solved before presenting the application on final demo day.

The user study focused on quantitative and qualitative analysis in order to answer following research question: *"How accurately and effectively does the 'Wish3D' mixed reality application convert voice commands into corresponding 3D models? How does the completion time of user interactions with 3D objects in the 'Wish3D' reflect on the system's usability?"*.

The quantitative methods focused on measuring the performance metrics of the application, providing clear insights into the functionality of the application. The qualitative measures explored the users' subjective experience with the application, focusing. The combination of these methods offers an extensive and well-rounded understanding of the application's real-world performance and user experience.

5.1. Tasks

The 5 tasks designed for the user study are closely linked to the research questions, aiming to evaluate the application's proficiency in converting spoken language into 3D models, and the effectiveness and efficiency of user interactions with these models. By including tasks such as generating and manipulating 3D objects based on voice input, the accuracy of the application's voice recognition as well that of the 3D model generation is directly assessed. Additionally, measuring the completion time for these tasks provides important insights into the responsiveness and efficiency of the system. This extensive approach addresses the impact of these attributes on the overall user experience with the application, by focusing on both of the key aspects: the technical performance and user interaction. Following are the tasks performed by users in User Study sequentially:

Task 1: *Open the App and go through the tutorial.*

Task 2: *Bring a monitor into the scene. Bring it close to you and then throw it into the distance.*

Task 3: *Bring an object of your choice into the scene and place it on a nearby table.*

Task 4: *Bring an object of your choice into the scene and rotate it.*

Task 5: *Bring an object of your choice into the scene and hit another existing object with it.*

5.2. Quantitative Measures of Wish3D

The quantitative measures focus on evaluating the performance and user experience of the application. For each task performed by the user, following metrics were noted down during user studies for quantitative analysis:

- **Task Completion Time:** Measures the time from voice command initiation to task completion.
- **Accuracy of Generated Object:** Users rate the accuracy of generated 3D objects from their commands.
- **Object Generation Time:** Time measured between given voice command and 3D object generation.
- **System Error Rate:** Frequency of errors.

5.3. Qualitative Measures of Wish3D

The qualitative assessment focused specifically on the *Wish3D* application itself and users' perception of it:

- **Ease of Use:** User feedback on their experience with the user interface, focusing on its intuitiveness and detail.
- **Quality of 3D Objects:** Users' perspectives on the generated 3D objects, including detail and prompt accuracy.
- **Immersive Experience:** User opinion on the immersive aspects of the experience, including the realism and natural behavior of objects, such as physics interactions.
- **Interaction with 3D Generated Objects:** User satisfaction related to interacting with the 3D objects, such as ease and intuitiveness of actions like rotating or throwing.

- **Voice Command Experience:** User feedback on using voice commands, specifically related to convenience and accuracy of command interpretation.

5.4. Formative and Summative User Study

Formative user studies were carried out on 6 users before the application was ready for the final demonstration. These users helped to identify problems with the application and potential areas for improvement. During Formative user study, users encountered following issues:

- The generated object was frequently placed behind the user, leading to instances where the user did not realise this and, consequently, did not look around to locate the object.
- Tags indicating the actions associated with buttons on the controller were often overlooked by users, hindering intuitive interaction.
- The duration for recording voice commands was sometimes too brief, resulting in users' commands being prematurely cut off.

Encountered issues were resolved before performing summative user study on final demo day on 10 new users.

5.5. Data Analysis

Data collected from User Studies were analysed to answer our research question.

Effectiveness is defined as the user's proficiency in task completion. In user studies, all participants accomplished the tasks, but with varying time durations. Users surpassing the upper bound of a box plot's distribution are considered not effective in that task. Efficiency is the amount of effort required to complete a task. It is calculated as the reciprocal of the average time taken to perform a task by a group. Data points representing ineffective performance are excluded from the efficiency calculation. We used Effectiveness and Efficiency as metrics to assess the interaction between users and *Wish3D*.

In order to answer the first research question i.e. *"How accurately and effectively does the 'Wish3D' mixed reality application convert voice commands into corresponding 3D models?"*, we plotted histogram of users rating on how accurate are the generated objects from voice commands (Figure 10). 3/5 was the most popular rating among users along different tasks. To test effectiveness of *Wish3D* object generation from voice command, we plotted box plots of object generation time for new vs experienced users as shown in Figure 11. *Wish3D* was effective 73 out of 80 times for new users and therefore holds 91.25 % effectiveness in generating 3D objects from voice commands.

In order to answer the second research question i.e. *"How does the completion time of user interactions with 3D objects in the 'Wish3D' reflect on the system's usability?"*, we first plotted box plots of completion time for each tasks for new vs experienced users (shown in Figure 12). It can be observed from Figure 12 that there is not much difference in median values of completion times for each tasks. To check for significance, we performed t-tests. P-value from t-tests are shown in Table 1. Since all the p-values are above 0.05, there is no significant difference between completion time of tasks for new users and experienced users. Therefore, we can say that *Wish3D* is a user-friendly application. We then calculated effectiveness of new users and efficiency of both new and experienced users. Effectiveness and Efficiency are shown in Table 1. It can be observed that *Wish3D* is more than 85% effective in all the tasks and new users are almost as efficient as experienced users except of Task 2 where efficiency of new users is much lower than experienced users. This drop can be explained by learning curve of users, since it was their first task after tutorial task (Task 1).

Task	**p-value**	**Effectiveness**	**Efficiency of new users (tasks/sec)**	**Efficiency of experienced users (tasks/sec)**
Task 1	0.4811	100.0	0.0128	0.0149
Task 2	0.1931	87.5	0.0197	0.0342
Task 3	0.3599	87.5	0.0268	0.0308
Task 4	0.2346	87.5	0.0319	0.0385
Task 5	0.2903	87.5	0.0314	0.0345

Table 1. Qualitative Analysis of Tasks

Error Category	**Count**
Premature Command Execution	2
Misinterpretation of Commands	3
User Interface Confusion	4
Object Recognition Issues	2
Interaction Issues	5

Table 2. Error Categories and Counts

We calculated System Error Rate as a quantitative measure in our user studies. Table 2 shows the System Error Categories and their corresponding counts observed during user studies. Our qualitative analysis suggests that users find controls/ interaction very intuitive. It was often brought up by users that the user interface and the interactions with the objects (i.e. rotating, pulling, pushing, throwing) are very instinctive. The overall user experience of the users was positive. However, some users shown concerns on lack of details in the generated 3D objects.

6. Results

In this section, we will present main results of *Wish3D* application and User Studies.

Table 3 shows the inference time of different models for a query "Chicken with a top hat". Inference time of 4 seconds of Shap-E model justifies our choice to integrate it in *Wish3D*.

Model Name	**Inference Time**
Shap-E	4s
One2345	45s
Zero123++, Gaussian Splatting	90s
SDXL	360s

Table 3. Inference time of different models for a query (chicken with a top hat) on a single RTX 3090

We recorded multiple timestamps during inference while doing the user studies, Table 4 contain averages of this data.

Step	**Time**
API overhead	1s
Speech to Text	4.5s
Text to 3d	4s
Mesh Processing	1s

Table 4. Approximate time spent on different steps of the inference pipeline when generating a 3d object from a recorded voice command.

Some main takeaways from users studies are: a) 3/5 is the most popular rating among users when asked about quality of generated 3D objects from voice commands. However, users were not satisfied with level of details in generated 3D objects b) *Wish3D* is 91.25% effective to generate 3D objects from Voice Commands c) New users were more than 85% effective given a task to perform in *Wish3D* d) New Users were almost equally efficient as Experienced users in all the Tasks except Task 2.

7. Discussions and Future Work

Based on user studies, *Wish3D* proves to be a user-friendly application, offering users the opportunity to express creativity and enjoy Mixed Reality. While new users may not have identified certain issues, experienced users have observed a notable concern – the absence of accurate mapping in specific regions. Although *Wish3D* incorporates Magic Leap's built-in mapping feature to enable interactions with the real world, its accuracy is inconsistent. Users have encountered problems such as objects falling through the floor, being thrown out of windows, and interacting poorly with reflective surfaces, influencing experience in certain situations.

The *Wish3D* application has potential for improvements: a) Integration of hand gestures can improve user experience to interact with 3D objects b) Granting users the option to choose from various 3D Generative Models. This would enable interested users to generate high-quality objects at the cost of longer inference times.

8. Conclusion

Our findings indicate that users perceive *Wish3D* as intuitive and interactive. Users were easily able to generate 3D objects of their choice and play around. However, there was an occasional dissatisfaction with the quality of the generated 3D objects. Shap-E produces 3D objects in real-time, providing users with real-time experiences, but at the expense of object quality. While other models excel in detailing the generated 3D objects, they have longer inference times. *Wish3D* is easily utilised by researchers investigating 3D object generation to assess the quality of their models by simply modifying the model in the backend.

References

[1] Mohamad Yahya Fekri Aladin and Ajune Wanis Ismail. Designing user interaction using gesture and speech for mixed reality interface. *International Journal of Innovative Computing*, 9(2), 2019. 1

[2] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions, 2023. 1, 2

[3] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering, 2023. 1, 2

[4] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization, 2023. 1, 2

[5] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. 1, 2

[6] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. 1, 2

[7] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model, 2023. 1

[8] Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Miguel Pino. fairseq S2T: fast speech-to-text modeling with fairseq. *CoRR*, abs/2010.05171, 2020. 2

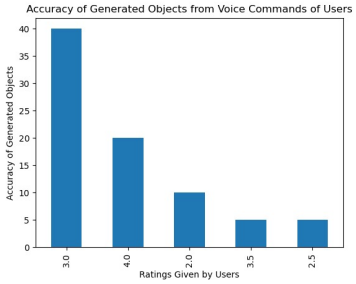


Figure 10. Accuracy of Generated Objects from Voice Commands (all users)

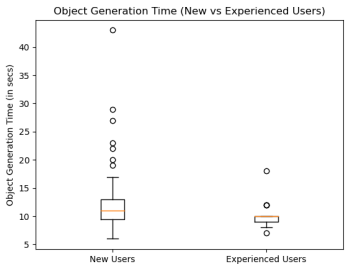
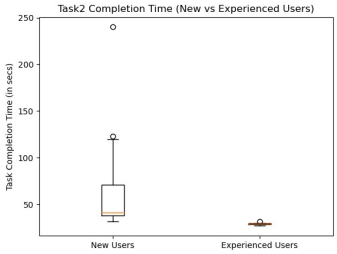


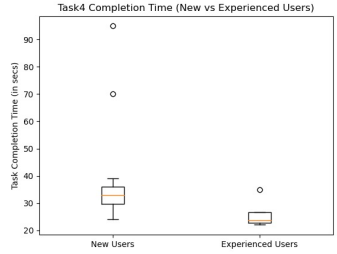
Figure 11. Object Generation Time for New Users and Experienced Users.

(a) Task 1

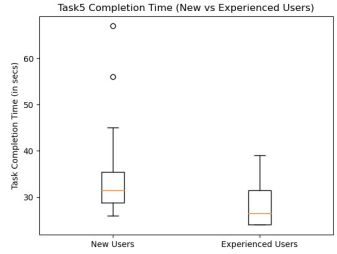


(b) Task 2

(c) Task 3



(d) Task 4



(e) Task 5

Figure 12. Completion Time for New vs Experienced Users