Towards Immersive Diminished Reality: Real-Time Object Removal Development and Evaluation of Visual Coherence using Vision-Language Models

Ryan Jay Chen

4/14/2025

Abstract

Diminished Reality (DR) techniques allow for the removal of real-world objects in augmented environments, offering a unique approach to modulating human perception through immersive spatial computing. This thesis proposes and evaluates a novel cross-platform DR pipeline designed to remove an object in real-time on head-mounted displays (HMDs) including the HoloLens2 and Apple Vision Pro (AVP). This thesis details an image processing pipeline, primarily built for the AVP, that integrates deep learning segmentation and inpainting (e.g. SAM, LaMa), optimized for on-device inference using Core ML. The work evaluates visual coherence of a sample image set created through the novel DR pipeline using vision-language models (VLMs) as a scalable proxy for human perception. This work concludes that real-time DR is feasible on current-generation hardware and holds promise for cognitive and perceptual interventions in industrial training, education, and therapeutic contexts.

1 Introduction

Augmented Reality (AR) technologies are primed to transform human interaction with physical reality by overlaying digital information onto the real-world environment. Diminished Reality (DR) exists as a subdomain of AR and involves the selective removal or suppression of visual elements. DR's unique ability to ”subtract” rather than ”add” to a user's vision introduces new potential for reducing distractions or improving situational awareness, thus providing applications in education, cognitive therapy, industrial safety, and immersive entertainment[1].

Recent hardware developments in head-mounted displays (HMDs) have enabled advanced spatial computing capabilities. Although Microsoft's HoloLens 2 (HL2) is no longer a cutting-edge device, it provides a robust platform for exploring the feasibility of DR techniques using Unity and the Mixed Reality Toolkit (MRTK). Many of its spatial mapping capabilities and its developer ecosystem are reflected in more recent HMDs such as the Apple Vision Pro

(AVP) and the Meta Quest Pro, making it ideal for prototyping. However, its optical see-through (OST) and additive-only rendering architecture limits its ability to convincingly remove objects from view—in optical see-through interfaces, darker objects appear more transparent, as beam-splitting does not effectively prevent real-world light from interfering with virtual images[2]. As such, it serves primarily as an initial testbed in this thesis.

In contrast, the AVP provides provides full video see-through (VST) capabilities, which enables direct manipulation of pixel content for more immersive DR applications. Although direct access to raw images is limited to enterprise licenses, a workaround to obtain raw images is provided via screen sharing. The AVP's hardware specifications also feature a powerful Neural Engine supporting real-time on-device ML inferencing, allowing for a more integrated and scalable image processing pipeline.

The central climax of this thesis is to build and evaluate an efficient real-time DR pipeline that can be deployed across platforms and adapted to their capabilities. For DR to be adopted in practice, especially on head-mounted displays (HMDs), its output must appear visually coherent and realistic. *Visual coherence* here refers to how seamlessly the modified region blends with the surrounding environment and appears plausible to the human eye.

Recent advances in VLM-based assessment inform of its novelty as a scalable, AI-driven alternative to traditional user studies. The VLM method allows DR-generated images to be scored on perceptual quality, such as blending, lighting, texture, occlusion, and artifact detection — evaluating DR through this lens provides both a practical benchmark and a stepping stone toward more immersive, real-time applications.

2 Literature Review

2.1 Pipeline Considerations for DR

DR techniques have roots in computer vision, image editing, and perceptual psychology. Earlier methods employed simple computer vision algorithms such as Telea and Navier-Stokes inpainting to fill regions of interest. Although these techniques are suitable for static or low-complexity scenes, they fail in dynamic, real-world settings.

Recent progress in deep learning has introduced state-of-the-art models like Large Mask Inpainting with Fourier Convolutions (LaMa) [3] and AOT-GAN [4], which can fill missing image regions with high semantic consistency, significantly improving upon classical inpainting techniques. Segmentation methods have also continually evolved, with tools like Segment Anything Model (SAM), FastSAM, and EdgeSAM enabling fast and flexible ROI extraction, even on-device [5, 6, 7].

The choice of hardware platform heavily influences DR strategy. HL2's additive rendering model prevents true occlusion of objects, making it ill-suited for realistic DR effects. Conversely, AVP's VST HMD architecture allows pixel-by-pixel control, enabling realistic object removal through compositing, a key requirement for immersive DR. Further, the AVP's improvements on spatial tracking features, such as responsiveness and tracking accuracy, in comparison to competitors in the field, prime it as a strong device for the heavy workload task of DR [8]. Moreover, Apple's Core ML pipeline provides a user friendly, comprehensive method of incorporating ML models in lightweight form for deployment on edge devices, such as iOS. The AVP is shown to have strong processing power built around ML processing compared to other HMDs, which makes it a suitable choice of HMD for Vision-heavy tasks [9].

2.2 Applications of DR and Influencing cognitive load

Literature across psychology and human-computer interaction further supports the utility of DR in managing attentional load. While traditionally, augmented reality systems improve task focus by overlaying helpful cues or information, thereby reducing cognitive load through guidance, DR takes the opposite approach yet aims for a similar outcome: subtracting irrelevant inputs to diminish distraction and enhance the user's cognitive performance. Studies show that distraction caused by visual or mobile stimuli—such as cell phones or unrelated content—can significantly impact attention span, learning outcomes, and information recall in both academic and real-world settings[10, 11, 12]. The concept of divided attention highlights that multitasking or visual clutter can deteriorate task performance, particularly under high perceptual load[13].

In addition, attentional tunneling, a phenomenon in which users fixate too strongly on virtual content to the detriment of surrounding physical awareness, has been documented in AR-supported learning environments[14]. While traditional AR may unintentionally exacerbate this effect, DR has the potential to reverse this by guiding attention toward relevant stimuli through saliency reduction. Visual distraction filtering and clutter removal, such as suppressing salient objects during search, have been shown to improve task efficiency and reduce error rates[15].

Along this vein, DR's strong potential application in the realm of cognitive load reduction has been demonstrated through recent experimental studies on cognitive distraction. For instance, Lee and Kim demonstrated that using an AR headset to "visually cancel out" a nearby smartphone (a common distracting object) significantly mitigates the cognitive distraction caused by its mere presence—with effects comparable to physically removing the phone from the room [16]. This finding underscores the potential of DR techniques to enhance focus: the AR-based removal of the phone led to improved performance on cognitive tasks, essentially freeing the user from the "brain drain" effect of an ever-present device.

Together, these studies point to the broader cognitive value of DR, namely, that reduced visual clutter and selective masking can support improved decision making, memory encoding, and task participation in information-rich environments.

2.3 Evaluation of AR Content by VLMs

While DR has been considered in prior work for cognitive modulation and distraction reduction, its success hinges on its perceptual realism—a factor often evaluated through subjective user studies. However, an emerging body of research shows that vision-language models (VLMs) can approximate human perception in AR image quality tasks. Duan et al. introduced the DiverseAR dataset, revealing that VLMs achieve up to 93% accuracy in perceiving and 71% in describing AR elements within images, highlighting their potential as proxies for human evaluation [17].

More relevantly, Duan et al. also compiled the ARQA dataset, composed of over 1,100 real-world and AR-sample image pairs from four AR platforms. Their work demonstrates how VLMs can rate images across specific visual factors: lighting direction and intensity, shadow realism, occlusion handling, and spatial plausibility [18]. Similarly, Itoh et al. describe visual coherence as a holistic measure that includes color and lighting consistency, texture continuity, artifact minimization, and object blending [19].

Although VLMs remain limited to static image interpretation, they represent a powerful tool for benchmarking DR realism. Future work may extend this approach to video-based assessment or real-time feedback loops for AR systems.

3 Methodology

3.1 System Architecture

The DR system, independent of hardware, consists of three main components: image acquisition, image processing, and image rendering. On the HL2, images are captured via the Photo-Video camera and sent to an edge server via TCP. On the AVP, two pipelines are outlined: image frames can be received over HTTP using a FastAPI-based pipeline for an edge-server based approach, or an on-device approach can be used via optimization of ML models using Core ML.

The HL2 was used primarily to test the feasibility of bounding box-based DR and back projection, though its additive-only rendering required visual workarounds that limited realism. These constraints motivated the transition to AVP, where VST allows greater visual flexibility. Once received, images are passed through a segmentation module (e.g., EdgeSAM) to identify objects for removal. The resulting mask is processed by an inpainting module (e.g., LaMa or AOT-GAN), and the completed image is sent back to the headset for rendering. The processed image is then used to texture either a flat surface (in the case of 2D projection) or a localized 3D mesh in the spatial environment. The latter provides stronger immersion, especially when combined with techniques like triplanar mapping using surface normals, which avoids stretching of textures that can occur with traditional UV mapping, particularly on irregular shapes [20].

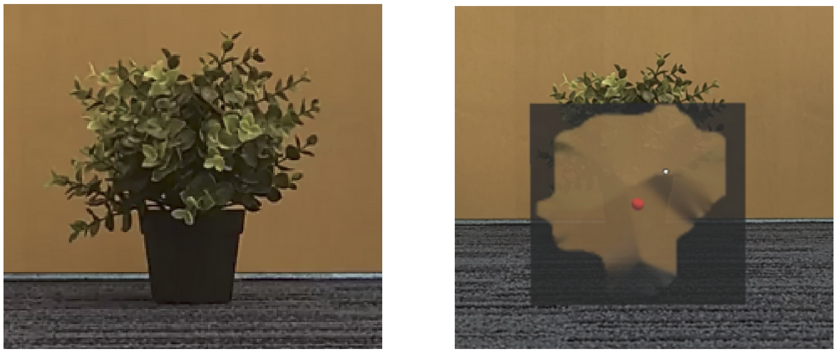
The general pipeline operates as follows:

1. **Capture:** Image frames are acquired from the main camera (AVP) or PV camera (HL2).

2. **Segmentation:** ROIs are identified using FastSAM, EdgeSAM, or U^2-Net.

3. **Inpainting:** Inpainting is performed using LaMa or AOT-GAN depending on available compute resources.

4. **Mesh Localization:** Bounding box or raycast methods are used to define a spatial region for the texture.



(a) Original Image (b) Telea Inpainted Image

Figure 1: Inpainting using edge-server in real-time on the HL2

5. **Rendering:** The processed image is mapped using projective or triplanar texture mapping

3.2 AVP-Specific On-Device Pipeline

This section proposes a real-time diminished reality (DR) pipeline for the Apple Vision Pro (AVP), focusing on image segmentation, object tracking, and deep inpainting—each optimized for the computational characteristics of head-mounted displays. The overall approach integrates Core ML, Metal shaders, and a photogrammetry-based object recognition pipeline to deliver an on-device experience with minimal latency [See Figure 4].

3.2.1 Image Segmentation with EdgeSAM

A central step in diminished reality is locating the object to be removed, which requires isolating a meaningful Region of Interest (ROI). This work adopts a point-prompt-based segmentation approach, wherein a single 2D point—generated either arbitrarily, by user input, or tracked object—is used to prompt a mask prediction. This segmentation task is handled using an adapted Core ML-converted EdgeSAM model that performs lightweight image segmentation on-device, returning a binary mask that identifies the object region to remove.

The output of this step is twofold:

1. The raw image (RGB input)

2. The masked image, in which the selected ROI is occluded using the binary mask.

These two images are passed as inputs to the inpainting model. This dual-input design ensures that the inpainting model has access to contextual scene information while clearly identifying the region to synthesize. The segmentation step is critical not only for quality but for computational efficiency, as smaller, well-targeted masks enable more localized and tractable inpainting inference.

Metal shaders are employed to preprocess and format the point prompt and image tensors efficiently, ensuring compatibility with the GPU-accelerated inference engine on the AVP.

3.2.2 Deep Inpainting with Core ML

The inpainting model used in this work is derived from the Big LaMa (Large Mask Fourier Inpainting) model[3], a state-of-the-art deep learning architecture known for high-quality reconstructions in occluded image regions. In its original form, Big LaMa contains over 51 million parameters and is too large for most edge devices. Using Core ML's quantization and compression pipelines, the model is successfully converted to a manageable 216.6MB Core ML-compatible format, enabling real-time use on AVP. Inference is handled entirely on-device, interfacing directly with the AVP's Neural Engine and GPU. By leveraging Apple's Metal Compute Pipeline, this implementation bypasses traditional CPU bottlenecks. A custom Metal shader is used to handle preprocessing (e.g., cropping, normalization, tiling), and output compositing is done efficiently in post-processing.

As inpainting is the most computationally expensive step of the pipeline, the system architecture is optimized to reduce the inpainted area as much as possible. Only masked regions are passed into the heavy inference step, enabling reduced latency and improved frame rates for real-time applications.

3.2.3 Object Tracking with Photogrammetry

Unlike traditional 2D object tracking pipelines (e.g., YOLO), this work utilizes a 3D photogrammetry based tracking system. A known object is scanned using multiple 2D images captured from varying angles. These images are processed into a high-fidelity `.usdz` 3D model using a photogrammetry toolchain, and the resulting model is used to generate a Core ML object classifier. At runtime, this classifier enables robust object recognition across varying viewpoints. Since the model has been trained on a complete 3D representation, it avoids the temporal lag that can occur when an object rotates or is seen from an uncommon angle. This ensures seamless performance during object manipulation in AR.

The primary limitation of this method is its reliance on pre-known objects. A setup phase is required where the object is captured and processed into a `.usdz` format. While this makes the system robust, it restricts spontaneity. Nonetheless, the object classifier model is lightweight enough for on-device inference and does not require server communication once loaded.

3.2.4 Core ML Integration and Metal Optimization

To ensure low-latency inference on-device, this work uses Apple's Core ML framework for both segmentation and inpainting. Pretrained models such as Big LaMa for large-mask inpainting are converted from PyTorch and TensorFlow formats into Core ML models. Quantization and model compression reduce the memory footprint significantly (e.g., Big LaMa is compressed to 216.6MB), facilitating deployment on AVP hardware. Metal shaders are employed to optimize both preprocessing (e.g., normalization, padding) and post-processing steps (e.g., compositing, overlay) [See Figure 3]. Inputs are fed into the GPU through a Metal

Figure 2: Photogrammetry scanned `.usdz` model of Charmander figure

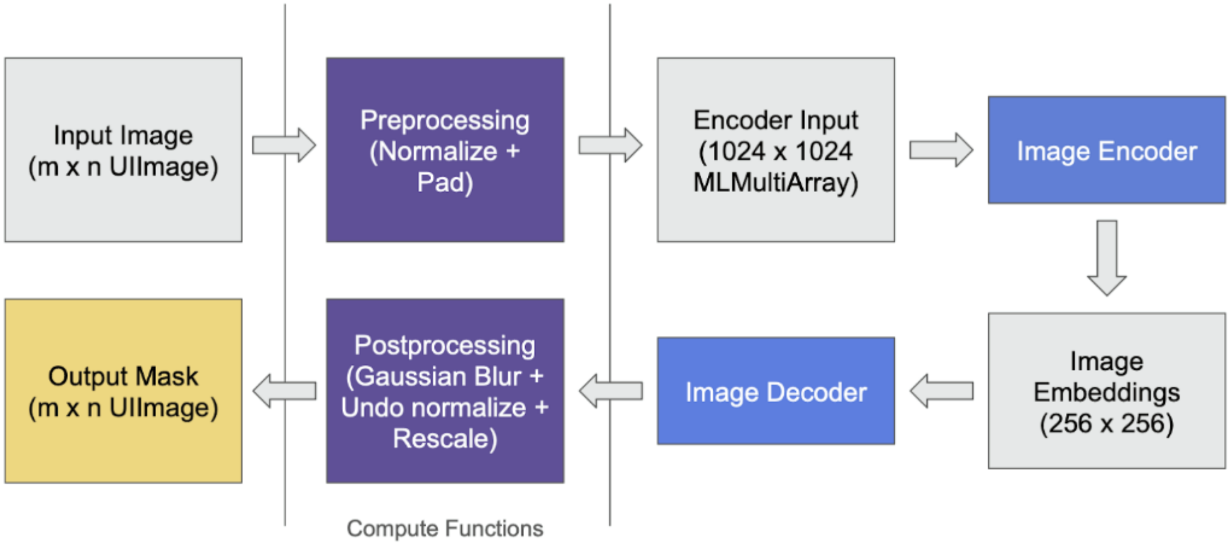


Figure 3: Metal Kernel Processing Pipeline for Machine Learning Models

Compute Pipeline that interfaces with Core ML's command buffer and encoder-decoder pipeline, taking advantage of AVP's Neural Engine and GPU for accelerated performance.

3.2.5 Raw Image Access on the AVP

One of the major challenges with the AVP is the lack of access to raw camera images without an enterprise developer license. To circumvent this, a screen sharing workaround is implemented. The AVP screen is streamed in real-time to a connected macOS device using the built-in screen sharing API. This image stream is captured on the Mac and sent to a Python-based TCP server, which acts as a proxy for raw image access on the AVP. The screen captured image is then relayed back to the AVP, for inpainting and segmentation task on-device.

This setup simulates an edge-server architecture but introduces noticeable latency, pri-

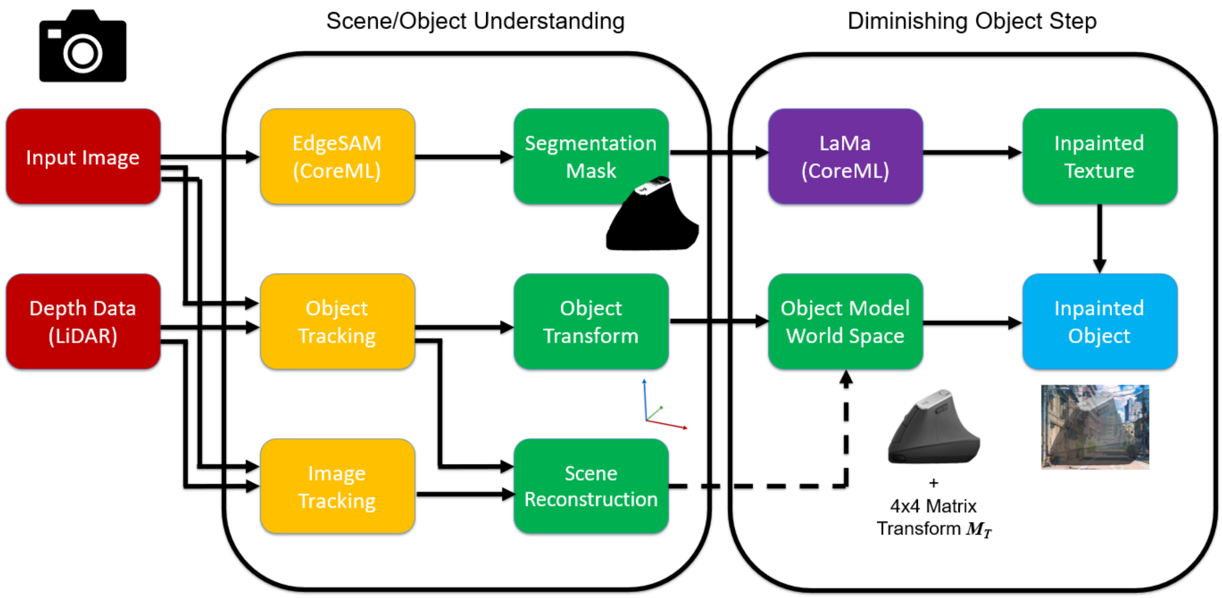
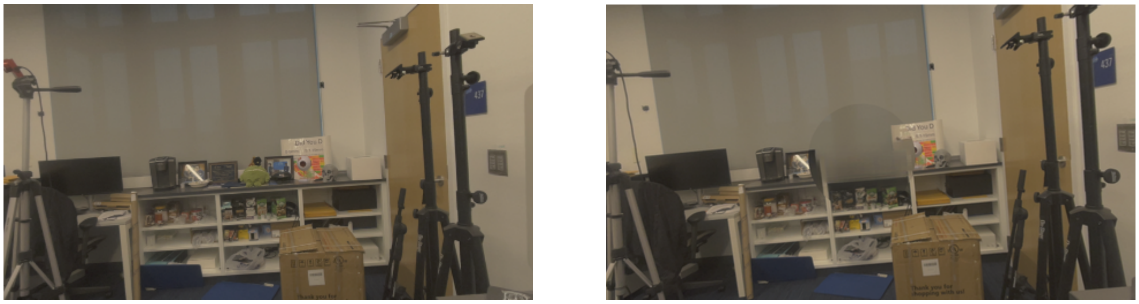


Figure 4: On-Device Diminished Reality Pipeline for AVP



(a) Original Image

(b) LaMa Inpainted Image

Figure 5: Inpainting with AVP On-Device Using LaMa

marily due to the round-trip overhead of capturing, transmitting, and re-rendering image data. Optimizing this architecture—by, for instance, embedding bidirectional communication within the screen sharing protocol or avoiding multiple port usage—represents an area of ongoing exploration. Furthermore, images received via screen sharing do not capture true raw data, but rather the pass-through augmented image data that includes AR holograms and models. As such, image processing may be done on a milliseconds previous image frame, depending on the latency/delay of frame transmission — which significantly hinders the usage of this pipeline.

4 Evaluation of Diminished Reality Visual Coherence by Vision Language Models

This thesis adopts the definition for *visual coherence* defined by Itoh et al. (Section 2.3), to guide DR evaluation. Rather than relying exclusively on subjective participant ratings or basic image similarity metrics, this work uses VLMs as a proxy for human/user evaluation of DR images. Duan et al. demonstrate that making prompts task-aware, without specifying AR or DR generated content removes cases of hallucination or bias from the VLM evaluation. Beyond this, when prompted to justify reasoning based on physical plausability in the scene, the VLMs accuracy is shown to improve [17]. VLMs are thus prompted with the following targeted question, which allows the VLM to focus on the most relevant perceptual features while abstracting away non-essential visual details:

"This is a real-world image in which one or more objects may have been digitally altered, reduced, or obscured. If you identify such an object, please explain how its removal affects the visual coherence of the scene at the location where the object was present. Consider how well that region integrates with its surroundings in terms of spatial alignment, lighting and shadows, texture and color blending, and the presence of artifacts or inconsistencies. First, rate the visual coherence of the edited region only, on a scale from 1 (poorly integrated) to 5 (perfectly seamless). Then, rate the overall visual coherence of the image, on a scale from 1 (poorly integrated) to 5 (perfectly seamless)."

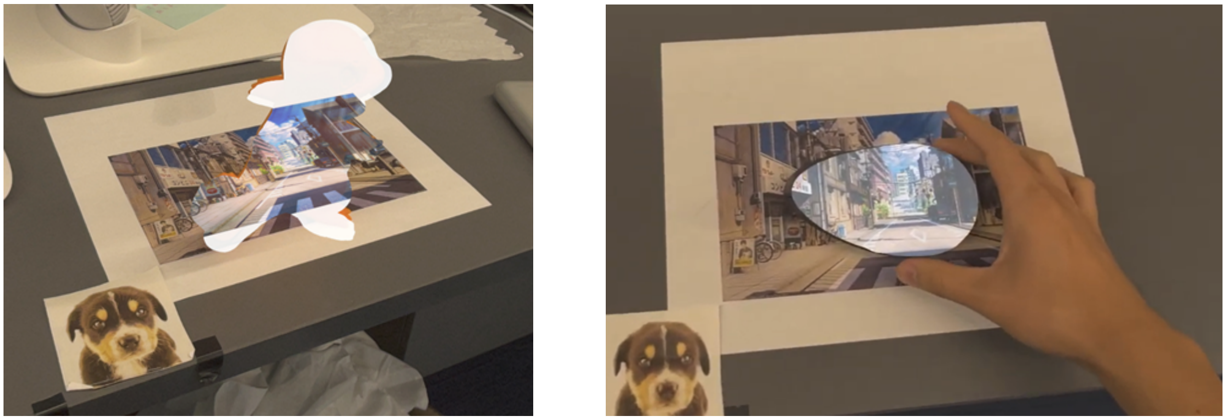
This prompt directs the model's attention toward the edited region and invites judgment across key criteria identified in prior work: color consistency, lighting accuracy, absence of artifacts, seamless blending, and plausible spatial geometry. These ratings serve as a proxy for visual realism in lieu of a full user study.

For image data collection, three different objects were used: an orange Charmander figurine, a Logitech computer mouse, and a Samsung Galaxy S25 phone. These models were scanned as `.usdz` objects via photogrammetry of image sets, either through the use of a curated image sequence or via the Reality Composer app on new-generation iPhones. Core ML reference objects were trained on each of the files and inputted into the app at runtime for object tracking. **98** images were collected through the DR pipeline on the AVP using the three objects (33 Charmander figurine photos, 20 Logitech mouse photos, 45 Samsung Galaxy S25 photos), with various different backgrounds and angles to represent a diverse range of possible DR environments. GPT-4o was prompted to provide ratings of visual coherence on a scale of 1-5. The results are aggregated in Section 5.3.

5 Results

5.1 HoloLens 2

The HL2 pipeline successfully demonstrated back-projected textures based on bounding boxes from image analysis. Frame updates occurred around every 500ms, which was accept-



(a) Charmander Figure

(b) Logitech Mouse

Figure 6: DR Prototype Image Results with Known Background

able for static or slowly changing scenes. However, hologram positioning was offset due to camera position relative to the eyes.

Nevertheless, HL2's additive-only rendering prevented true occlusion, limiting the DR illusion - any darker background surfaces appeared transparent or translucent, exposing the object to be diminished. These findings validated HL2's role as a prototyping platform but underscored the need for a VST device for compelling DR.

5.2 Apple Vision Pro

The AVP's VST architecture allowed for more immersive DR via object tracking, fast and reactive image processing, and pixel editing capability. With Core ML—LaMa and EdgeSAM running on-device, latency was minimal. Real-time projection of inpainted regions onto spatial meshes was achieved with high visual fidelity.[1] [2]

Inference times were reduced to below 100ms per frame for segmentation and inpainting. Visual alignment remained an issue in stereo view but was mostly resolved using projection correction algorithms. Visual pose delay was also observed due to the smooth lerp feature of the model's movement, but did not affect the functionality of the app.

5.3 Results from VLM Evaluation

The visual coherence of three digitally removed objects — a Charmander figurine, a computer mouse, and a Samsung Galaxy phone — was evaluated across a range of real-world backgrounds. Two visual coherence metrics were assessed using a vision-language model: overall image coherence and edited region coherence. In three instances involving the Samsung Galaxy phone, the VLM did not detect any significant visual anomalies indicative of object removal. Edited region coherence scores ranged from 2 to 5, while overall image

[1] Demonstration videos of the concept are found here: https://duke.is/drvideo1

[2] https://duke.is/drvideo2

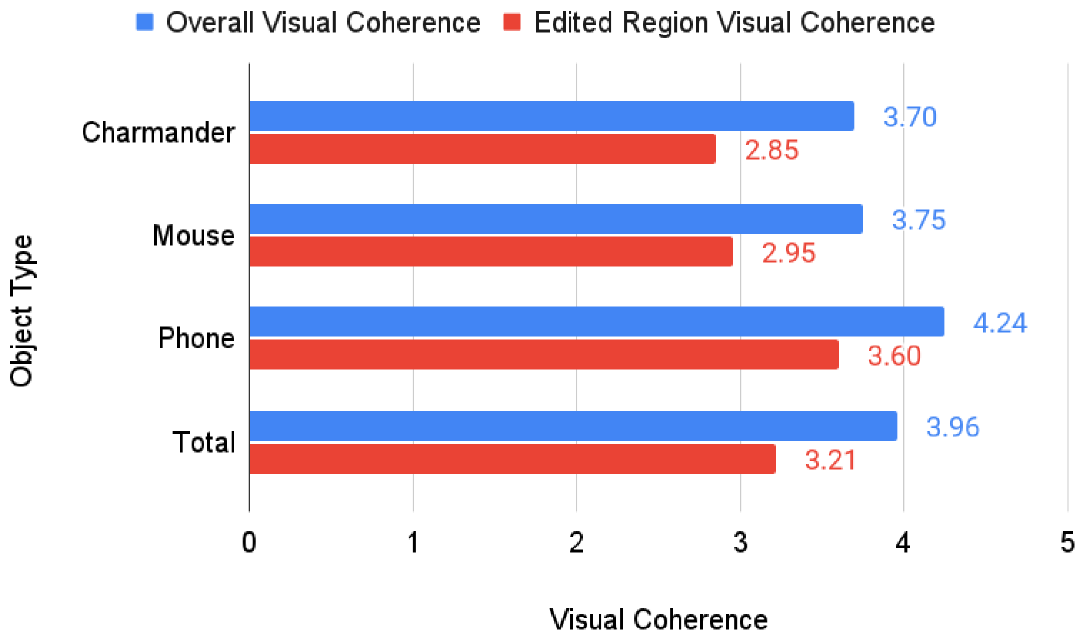


Figure 7: Visual Coherence of DR Images

coherence scores ranged from 3 to 5. Notably, several images were assigned the maximum coherence score, suggesting that the visual edits were judged to be indistinguishable from unaltered imagery (Figure 9). Conversely, other scenes exhibited lower coherence ratings, reflecting noticeable disruptions or artifacts following object removal (Figure 10).

As shown in Figure 7, the Samsung Galaxy phone model achieved the highest average edited region coherence and overall visual coherence (3.60 and 4.24, respectively), followed by the mouse (2.95 and 3.75), and the Charmander figure (2.85 and 3.70). A similar trend was observed in overall coherence, though all objects scored higher with the overall coherence score than edited region score, suggesting that there are some improvements that can be made to the existing pipeline to improve upon local visual artifacts in scene.

6 Discussion

This thesis proposes and evaluates a novel DR pipeline using visual coherence as a core metric for effectiveness of the pipeline. HL2 served as an effective tool for early feasibility testing of DR pipelines, although its optical see-through design inherently restricted the illusion of object removal. In contrast, AVP's VST display and compute capacity demonstrated greater suitability for DR.

Evaluation by GPT-4o of the DR-generated scene images reveals that on-device diminishing reality is highly feasible, and can be accomplished in real-time given a few optimizations to existing computer vision architectures. While overall visual coherence with this diminished reality pipeline remains relatively high (¿3), there is substantial room for improvement in the local visual coherence of digitally reduced objects given an average score ¡ 3. Notably, the lowest overall coherence score observed was 3, suggesting that in holistic views of the scene, diminished objects do not significantly disrupt the visual realism of the scene and are not overly conspicuous to the viewer. This finding implies that the DR system achieves a level of visual plausibility comparable to typical AR hologram imagery through the lens of a VLM.

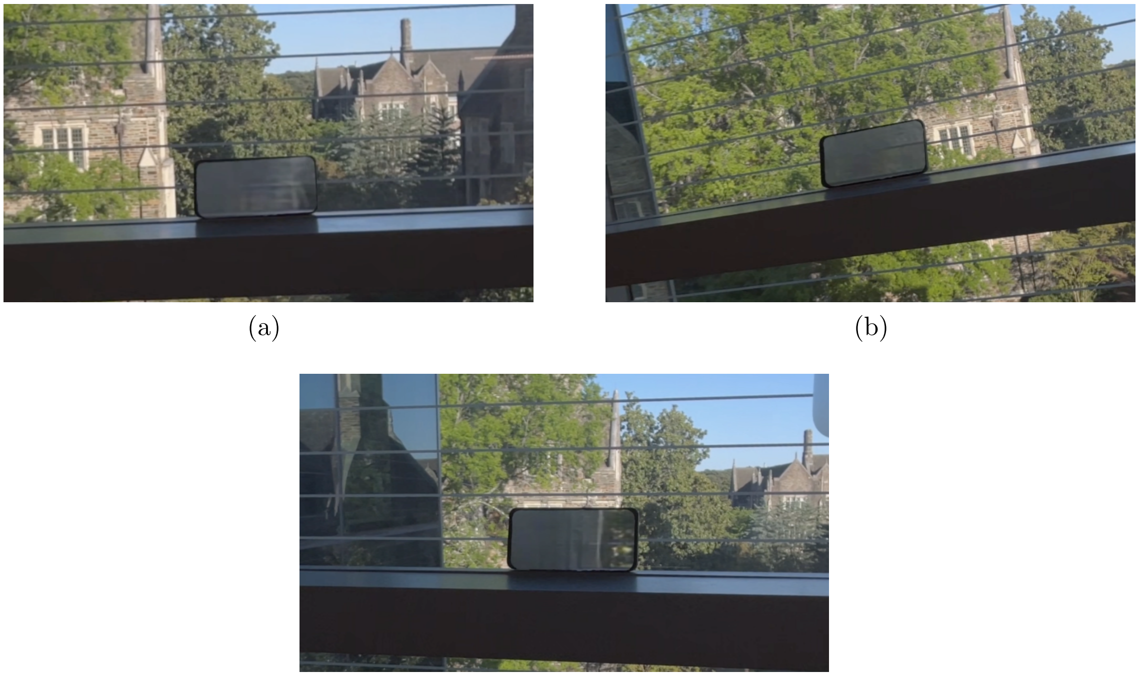
Still, edited region coherence scores reveal nuanced challenges. For example, objects with complex geometry often received lower coherence scores (typically rated a score of 2 or 3), corresponding to misalignments between the inpainted textured mesh and underlying object geometries, especially in the case of the Charmander figurine and the Logitech mouse. These types of objects possess irregular surfaces with high vertex density, presenting challenges towards the object tracking and segmentation processes of the pipeline. Inaccurate object masks then lead to visible edge artifacts, poor alignment, and texture bleeding, all of which degrade local visual realism and coherence.

In some instances, low coherence ratings were not influenced by the diminished object, but rather by visual elements related to the grounding or UI entities of the AVP app — for example, shadows cast by the application window or automatic adjustments to lighting. This introduces a confounding factor, in that it becomes difficult to isolate whether low visual coherence is attributable to the DR pipeline itself, or extraneous visual cues unrelated to the object removal process. Future evaluation frameworks may benefit from explicitly decoupling these elements to isolate coherence issues.

The image results involving the Samsung Galaxy phone were particularly promising. The phone achieved consistently high average coherence scores above 3, with several scenarios evaluated at the maximum rating of 5, and others not even recognized as digitally altered. These results signify the ability of the DR pipeline to handle rectangular, well-defined objects and suggest that inpainting-based diminishing texture synthesis is convincing when combined with effective object tracking and segmentation.

However, another confounding factor may lie in the ambiguities in VLM perception of the phone. Given its natural rectangular form and display screen, a partially masked or imperfectly inpainted phone may still be interpreted by the VLM as a real phone screen, rather than an altered region. This suggests that VLMs may be biased towards interpreting certain geometries (such as rectangles) as plausible by default, even in the presence of subtle inconsistencies. Further investigation into misguiding VLMs is needed, particularly in evaluating the reliability of VLM-based image quality and realism assessment for AR and DR tasks.

In summary, these findings underscore the dynamic interplay between object complexity, background context, VLM prompting, and the DR pipeline's performance in determining the visual plausibility of diminished reality. While the system performs well under controlled conditions, future work will need to address and improve upon segmentation robustness, object tracking mesh alignment, and strategies for decoupling coherence artifacts introduced by non-DR elements such as external UI.

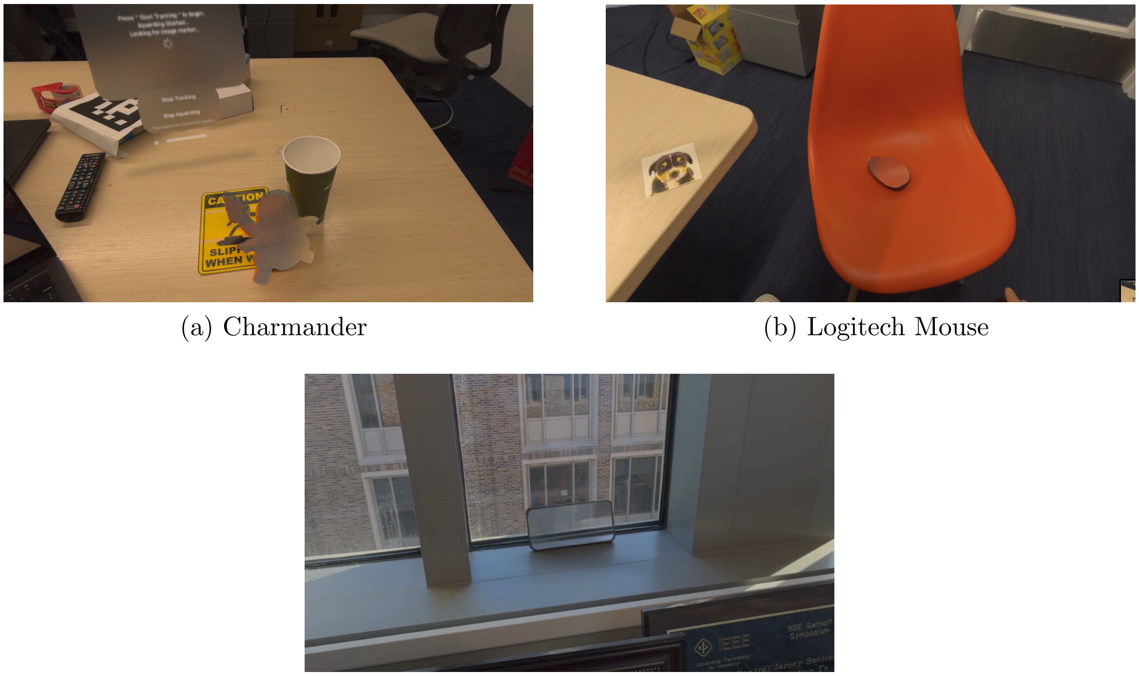


(a)

(b)

(c)

Figure 8: Images with No Diminished Reality Detected

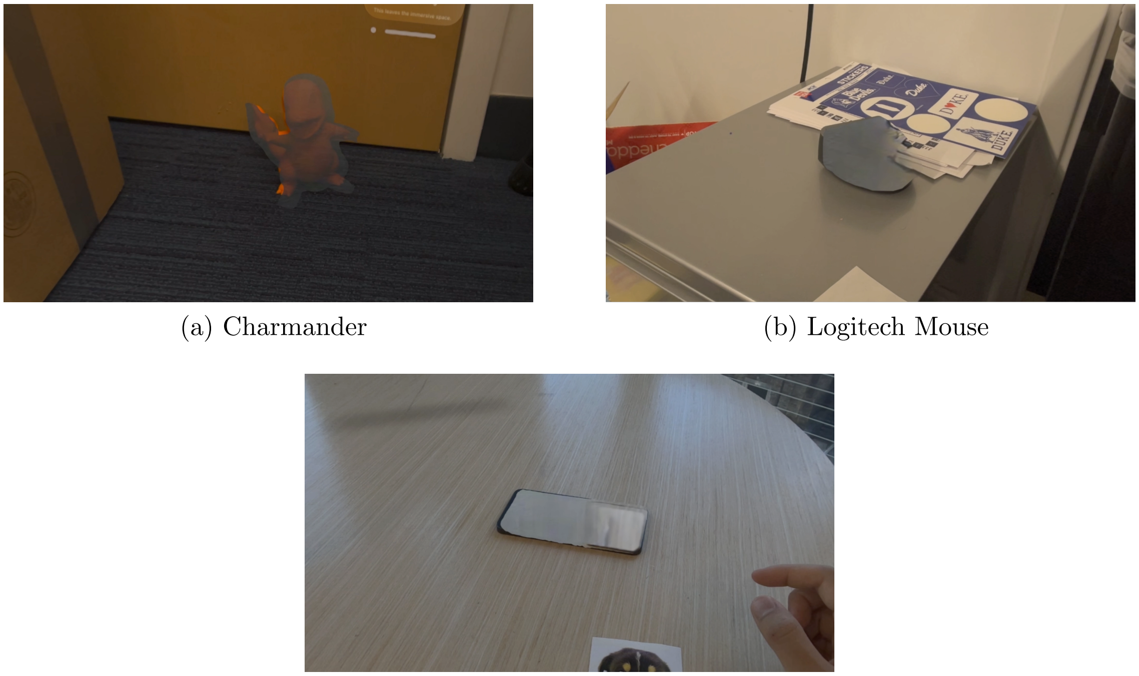


(a) Charmander

(b) Logitech Mouse

(c) Samsung Phone

Figure 9: Images with Edited Region Visual Coherence Scores of 5



(a) Charmander

(b) Logitech Mouse

(c) Samsung Phone

Figure 10: Images with Edited Region Visual Coherence Scores of 2

6.1 Limitations and Future Work

This research is primarily focused on the design and technical implementation of a diminished reality pipeline for the Apple Vision Pro, with emphasis on demonstrating feasibility and performance under constrained hardware and software conditions. While the current pipeline effectively demonstrates diminished reality on the Apple Vision Pro, several areas remain open for enhancement.

Object Tracking A key direction is enabling real-time 3D object learning to replace the current reliance on pre-scanned `.usdz` models. Implementing a brief setup phase where the user captures multiple views of an object could allow for spontaneous integration of new items into the tracking system. This would significantly broaden the usability of the pipeline, especially in dynamic or uncontrolled environments. In parallel, localized and hierarchical inpainting strategies could be implemented to reduce the inference load. Rather than processing the entire image, future models could focus on the masked region with sparse attention mechanisms or adopt a multi-stage refinement process, improving runtime without sacrificing visual quality.

Model Performance Improvements Additional performance improvements can be achieved through deeper integration with the Metal Compute Pipeline. Techniques such as adaptive tiling, asynchronous command encoding, and model pruning could further optimize GPU utilization on the AVP. This would open the door to supporting more advanced models, including diffusion-based inpainting or hybrid semantic-driven approaches, provided they are quantized effectively for edge deployment, as well as motion-aware or kinetic DR systems for dynamic objects and environments. Moreover, expanding the model ecosystem beyond EdgeSAM and LaMa—for example, using lightweight MobileSAM variants or incorporating zero-shot learning methods—would allow for greater flexibility across diverse scenes and objects.

Screen-Share Latency Addressing the latency introduced by screen-sharing workarounds is essential for closing the gap between research prototypes and real-time DR applications. Future work will explore optimizing the current image relay mechanism by embedding bidirectional communication within the existing screen-sharing protocol, or utilizing shared memory to eliminate redundant data transfer steps.

VLM Evaluation Improvements One major limitation of VLM evaluation was the lack of background or raw images for comparison with the DR-altered images. Without a reference image to compare, the VLMs understanding is lesser compared to previous similar studies on VLM image quality assessment of AR scenarios. In the future, DR work on the AVP should use background and DR image pairs to evaluate the scenee for a more holistic set of modalities.

Training VLMs specifically for tasks in DR for higher fidelity, greater image quality assessment, and more robust scene understanding capabilities would enable the integration of domain-adapted perceptual systems into real-time AR pipelines. For example, models could be fine-tuned on DR-specific datasets that include examples of object removal, inpainting artifacts, and varying degrees of coherence. With improved visual sensitivity to issues like visual coherence artifacts, task-specialized VLMs could act as real-time evaluators—providing automated feedback during the development of DR applications. In doing so, this integration would facilitate feedback-loop based optimization, allowing DR systems to dynamically adjust processing methods to maximize perceptual realism and user quality of immersion and quality of experience across diverse environments.

Future work will incorporate systematic benchmarking and user-centered evaluations to better quantify performance, latency, perceptual realism, and potential cognitive impacts of diminished reality in various environments. This initial implementation lays the necessary foundation for such studies by validating the feasibility of the technical pipeline under realistic constraints.

Altogether, these enhancements aim to make the DR experience more adaptive, efficient, and deployable at scale, while providing tools for assessment and streamlining in the future.

7 Conclusion

This thesis presents a practical framework for building real-time diminished reality systems on modern head-mounted displays, with a particular focus on development for the Apple Vision Pro. By integrating image segmentation, deep inpainting, and spatial rendering into a unified pipeline, this work demonstrates that DR is not only technically feasible on HMDs but can also be implemented efficiently through optimized on-device inference and GPU acceleration.

Rather than focusing solely on application-level metrics like task performance or cognitive load, this work prioritizes perceptual realism as a foundational step towards application.

VLM-based evaluation represents a promising proxy for human perception and sets the stage for iterative improvements in DR rendering.

As DR transitions toward mainstream usage, tools that measure and improve visual coherence will be essential. This work contributes a framework for doing so and identifies future directions for kinetic DR, dynamic scene evaluation, and hybrid human-AI assessment, which highlight the potential DR has to reshape interaction paradigms by selectively removing information to improve focus, reduce distraction, and enhance task-oriented environments.

References

[1] Y. F. Cheng, H. Yin, Y. Yan, J. Gugenheimer, and D. Lindlbauer, “Towards understanding diminished reality,” in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI ’22, (New York, NY, USA), Association for Computing Machinery, 2022.

[2] K. Kim, A. Erickson, A. Lambert, G. Bruder, and G. Welch, “Effects of dark mode on visual fatigue and acuity in optical see-through head-mounted displays,” in *Symposium on Spatial User Interaction*, SUI ’19, (New York, NY, USA), Association for Computing Machinery, 2019.

[3] R. Suvorov *et al.*, “Resolution-robust large mask inpainting with fourier convolutions,” *arXiv preprint*, vol. arXiv:2109.07161, 2021.

[4] Y. Zeng *et al.*, “Aggregated contextual transformations for high-resolution image inpainting,” *arXiv*, 2020.

[5] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, “Segment anything,” 2023.

[6] X. Zhao, W. Ding, Y. An, Y. Du, T. Yu, M. Li, M. Tang, and J. Wang, “Fast segment anything,” 2023.

[7] C. Zhou, X. Li, C. C. Loy, and B. Dai, “Edgesam: Prompt-in-the-loop distillation for on-device deployment of sam,” 2024.

[8] T. Hu, F. Yang, T. Scargill, and M. Gorlatova, “Apple v.s. meta: A comparative study on spatial tracking in sota xr headsets,” in *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, ACM MobiCom ’24, (New York, NY, USA), p. 2120–2127, Association for Computing Machinery, 2024.

[9] D. Khan, X. Liu, O. Mena, D. Jia, A. Kouyoumdjian, and I. Viola, “Loxr: Performance evaluation of locally executing llms on xr devices,” 2025.

[10] F. Sana, T. Weston, and N. J. Cepeda, “Laptop multitasking hinders classroom learning for both users and nearby peers,” *Computers & Education*, vol. 62, pp. 24–31, 2013.

[11] B. Thornton, A. Faires, M. Robbins, and E. Rollins, "The mere presence of a cell phone may be distracting: Implications for attention and task performance," *Social Psychology*, vol. 45, no. 6, pp. 479–488, 2014.

[12] A. F. Ward, K. Duke, A. Gneezy, and M. W. Bos, "Brain drain: The mere presence of one's own smartphone reduces available cognitive capacity," *Journal of the Association for Consumer Research*, vol. 2, no. 2, pp. 140–154, 2017.

[13] M. S. Cain and S. R. Mitroff, "Distractor filtering in media multitaskers," *Attention, Perception, & Psychophysics*, vol. 73, pp. 1633–1641, 2011.

[14] J. L. Gabbard *et al.*, "Attentional tunneling and ar design: Guidelines for creating effective ar experiences," in *Proceedings of CHI 2021*, pp. 1–12, 2021.

[15] N. Gaspelin and S. J. Luck, "Suppression of salient objects prevents distraction in visual search," *Journal of Neuroscience*, vol. 38, no. 32, pp. 7843–7853, 2018.

[16] J. Lee and L. H. Kim, "Diminishar: Diminishing visual distractions via holographic ar displays," 2025.

[17] L. Duan, Y. Xiu, and M. Gorlatova, "Advancing the understanding and evaluation of ar-generated scenes: When vision-language models shine and stumble," 2025.

[18] L. Duan, Y. Xiu, S. Eom, R. J. Chen, C. Li, Y. Hu, and M. Gorlatova, "Bridging human perception and automated evaluation: Vision-language model-based visual quality assessment of ar-generated scenes." Submitted to IEEE International Symposium on Mixed and Augmented Reality, 2025.

[19] Y. Itoh, T. Langlotz, J. Sutton, and A. Plopski, "Towards indistinguishable augmented reality: A survey on optical see-through head-mounted displays," *ACM Comput. Surv.*, vol. 54, July 2021.

[20] K. Schuster, P. Trettner, P. Schmitz, and L. Kobbelt, "A three-level approach to texture mapping and synthesis on 3d surfaces.," *Proc. ACM Comput. Graph. Interact. Tech.*, vol. 3, no. 2, pp. 1–1, 2020.