2026 European Conference on Computing in Construction
Corfu, Greece
July 12–15, 2026

Text-to-AR for Home Remodeling: An End-to-End Pipeline from Natural Language to AR-Compatible 3D Assets

Vamsi Sai Kalasapudi[1], Bharadwaj R. K. Mantha[2,3], Hassan Odeh[4] and Alekhya Seelam[1]
[1]College of Computing, Engineering and Construction, University of North Florida, Florida, USA
[2]Dept. of Civil and Environmental Engineering, University of Sharjah, Sharjah, UAE
[3]Dept. of Civil Engineering, Indian Institute of Technology Madras (IITM), Chennai, India
[4]Freelance AI Consultant, UAE

Abstract

Home remodeling often relies on sketches and informal discussions, causing miscommunication, rework, and added cost. Existing AR tools are typically proprietary and limited to predefined 3D asset libraries. This study develops and validates an executable end-to-end pipeline that converts free-form homeowner text into AR-compatible 3D remodeling visualizations. A sofa case study evaluates generated assets across image resolution, input type, and quality condition using SSIM and dimensional metrics. Results show visually coherent assets suitable for early-stage communication, while highlighting depth and scale limitations for future environment-aware AR remodeling tools.

Introduction

The home remodeling market has expanded significantly in the post-pandemic period, exceeding $880 billion globally in 2022 and projected to surpass $1.1 trillion by 2027, (ReportLinker, 2023). The United States alone accounted for $567 billion in 2022 (Joint Center for Housing Studies of Harvard University, 2023). This growth has accelerated digital and remote planning workflows, with increasing reliance on immersive tools such as VR and AR, now accessible via consumer devices (Brown, 2020). Augmented Reality (AR) overlays digital elements onto real-world environments, enabling users to visualize furnishings and design changes within their own spaces before implementation (Azuma, 1997; IKEA, 2017; Inc., 2018).

Consumer applications such as IKEA Place and Houzz's "View in My Room 3D" demonstrate the growing adoption of AR in residential design. These tools improve decision-making and purchasing likelihood, with Houzz reporting significantly higher engagement and conversion rates among AR users (IKEA, 2017; Inc., 2018). However, these systems rely on proprietary, pre-modeled catalogs, limiting scalability, customization, and the ability to handle free-form user inputs (e.g., "two-seater U-shaped red sofa") (Boopathy et al., 2024).

AR is also being explored in construction workflows, such as the Augmented Carpentry project, which overlays fabrication guidance directly onto materials, demonstrating integration of digital design with physical execution (Coxworth, 2025). This AR-assisted production method adeptly merges human skill with computational precision, actualizing digital designs on the workpiece. These advancements indicate a wider trend of employing AR to seamlessly incorporate design information into actual construction processes, extending beyond mere screen visualization. Despite these advances, remodeling remains communication-intensive, particularly for homeowners with limited technical expertise. Reliance on informal sketches and verbal descriptions often leads to misinterpretation, delays, and increased costs (for Housing Studies of Harvard University, 2014). With only basic visualization support, homeowners often with little to no construction experience, frequently rely on hand-drawn sketches from imagination and informal discussions at home to explain their ideas. These ad hoc methods can make it difficult to convince the remodeler of the intended design, increasing the likelihood of miscommunication, design inaccuracies, and inefficient implementations. Many existing tools assume user familiarity with design principles or rely on fixed 3D asset libraries, limiting creativity and personalization. Even AI-based systems typically provide predefined options rather than enabling translation of nuanced, real-world inspirations into customized designs. Furthermore, end-to-end pipelines converting free-form input into AR visualizations remain largely unavailable (Höllein et al., 2023).

To address these gaps, this study presents a first-step toward automating interior remodeling visualization by converting free-form user descriptions into AR-compatible 3D models. The goal is to reduce communication barriers between non-expert homeowners and professional stakeholders including contractors, sub contractors, suppliers, interior designers, and architects. The proposed system enables users to describe desired designs in natural language and generates corresponding AR visualizations that are projected onto the real world environments. It integrates natural language processing (NLP), text-to-image generation, multi-view synthesis, 3D reconstruction, and AR deployment into a unified pipeline, generating content dynamically rather than relying on pre-built assets. The aims of this research are to (1) streamline and accelerate early-stage ideation by producing realistic previews from textual or image descriptions, (2) reduce the cost and specialized expertise typically required to create custom 3D remodeling visualizations, and (3) demonstrate the feasibility of end-to-end automation from text or image input to AR output.

Literature Review

Advances in generative AI, computer vision, and AR have expanded automated content creation in interior design. This review synthesizes progress in text-to-image, multi-view synthesis, and text-to-3D reconstruction, highlighting limitations in achieving end-to-end AR workflows for homeowner use. Diffusion-based models such as Stable Diffusion (Rombach et al., 2022) and Imagen (Saharia et al., 2022) generate high-quality images from text, with domain adaptations enabling interior design visualization (Chen et al., 2023). However, outputs remain 2D and unsuitable for AR interaction.

Multi-view synthesis methods such as SceneScape (Fridman et al., 2023) and FlexGen (Zhang and Lee, 2023b) improve geometric consistency across viewpoints. However, they remain computationally intensive and prone to inconsistencies, with limited support for iterative refinement. 3D reconstruction methods such as NeRF (Mildenhall et al., 2020b), Text2Room (Höllein et al., 2023), DreamFusion (Poole et al., 2022), and Magic3D (Lin et al., 2023) enable text-driven 3D generation. However, outputs typically require extensive post-processing (e.g., mesh cleanup, texturing), limiting usability for non-experts. A key limitation is AR integration, as generated 3D assets often require manual conversion for real-time deployment. Existing research pipelines rarely support direct AR-ready outputs, limiting scalability (Poole et al., 2022; Lin et al., 2023).

The proposed pipeline addresses this gap by integrating text input, multi-view synthesis, 3D reconstruction, and direct export to AR-compatible formats (e.g., USDZ), enabling iterative and scalable remodeling visualization. Table 1 summarizes how representative recent studies cover the major components required for a text-to-AR workflow. Prior work shows strong progress in individual stages such as text-to-image generation, multi-view synthesis, and text-to-3D reconstruction. However, most approaches do not provide a complete path to AR-ready deployment and often offer limited support for iterative refinement that matches real remodeling conversations.

Table 1: Comparative analysis of recent text-to-AR methodologies

Study	Text-to-Image	Multi-view Synthesis	3D Reconstruction	AR Integration	Iterative Refinement
(Chen et al., 2023)	Yes	No	No	No	Yes
(Fridman et al., 2023)	Yes	Yes	Yes	No	Limited
(Höllein et al., 2023)	Yes	Yes	Yes	No	No
DreamFusion (Poole et al., 2022)	No	No	Yes	No	No
Magic3D (Lin et al., 2023)	No	No	Yes	No	No
Our Proposed Pipeline	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

Reliable 3D Reconstruction as a Practical Requirement

Reliable 3D reconstruction is essential in remodeling, as decisions depend on spatial accuracy and scale rather than visual plausibility alone. Single-image reconstruction methods suffer from missing geometry due to occlusions (Mildenhall et al., 2020a; Gkioxari and Malik, 2019), while traditional multi-view approaches require extensive image capture (Seitz and Curless, 2021; Schönberger and Frahm, 2016). Generative multi-view synthesis offers a promising alternative by producing consistent viewpoints from text, reducing data requirements while improving reconstruction fidelity (Ho and Salimans, 2023; Zhang and Lee, 2023a). These observations motivate the need for workflows that couple multi-view consistency with efficient reconstruction and deployment constraints, aligning with the introduction's goal of advancing toward automated transformation from free-form user intent into AR-compatible 3D assets for remodeling visualization.

Methodology

This study presents a first-step exploration toward an end-to-end text-to-AR workflow for home remodeling visualization. The goal is not to introduce new generative or reconstruction models, but to establish an executable baseline and clarify what is practically required to translate a homeowner's free-form natural language intent into an AR-ready 3D asset that can be viewed within the real home environment. The contribution of this work is therefore the integration of established components into a deployable workflow with inspectable intermediate outputs, which supports intent alignment and helps identify where failures and quality limitations arise in practice. In this context, the pipeline is designed to reduce the homeowner–professional communication gap by turning ambiguous verbal descriptions into shared, inspectable visual artifacts that can be iteratively refined during early-stage remodeling discussions. This methodology aims to establish an end-to-end workflow that transforms free-form natural language descriptions into AR-ready 3D assets for early-stage home remodeling visualization. It further seeks to characterize how practical input conditions and key pipeline settings influence output quality, particularly reconstruction robustness and dimensional plausibility. Finally, the workflow introduces intermediate quality checks to limit error propagation and documents where human refinement remains necessary to resolve intent ambiguity and support clearer homeowner–professional communication.

Pipeline Overview

The workflow consists of four interrelated phases that progressively convert language into deployable AR content: (1) Text-to-2D Synthesis, (2) Image Enhancement and Multiview Generation, (3) 2D-to-3D Reconstruction, and (4) 3D-to-AR Integration. Each phase produces a tangible intermediate output (candidate images, processed multiview sets, reconstructed meshes, AR-compatible assets) that can be inspected and improved before moving forward, which supports faster alignment between user intent and the generated visualization. This staged design is used as a practical first step because it produces inspectable intermediate outputs and an explicit mesh for AR deployment, which is important for maintaining geometric consistency and dimensionally plausible visualization in remodeling contexts.

Figure 1: Overview of the proposed Text-to-AR pipeline. The shaded components indicate the phases evaluated in the case study.

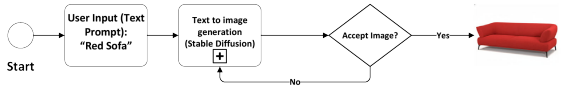

Figure 2: Phase 1: Text-to-2D synthesis workflow.

Phase 1: Text-to-2D Synthesis

Phase 1 translates a user's natural language description (e.g., "modern red two-seater sofa") into an initial set of candidate images that serve as the visual starting point for later multiview generation and reconstruction. Stable Diffusion is used as the text-to-image engine, and multiple candidates are produced per prompt by varying seed values and guidance settings to capture reasonable visual interpretations of the same intent. In this study, a Stable Diffusion model (v1.5) is used to generate candidate images, with multiple outputs produced by varying random seed values to support iterative selection. To directly support remodeling conversations, this phase is treated as an iterative intent-clarification step. Rather than assuming the first output is correct, users can select the closest candidate and refine the description to better match their target appearance. This human-in-the-loop selection is intentionally positioned as a lightweight communication mechanism: it reduces ambiguity early, before downstream 3D reconstruction amplifies errors or inconsistencies.

The current implementation does not explicitly encode metric dimensions in text prompts, as diffusion generation operates in a scale agnostic space. However, preliminary tests with dimension aware prompts (e.g., approximate size or capacity) showed improved proportional consistency.

Phase 2: Image Enhancement and Multiview Generation

After Phase 1, the selected image is conditioned to better support downstream 3D reconstruction. This step focuses on improving visual clarity and isolating the object so that later multiview synthesis and reconstruction are less sensitive to background clutter and image artifacts. Multiview synthesis is implemented using Zero123++ (sudo-ai/zero123plus-v1.2) with an Euler Ancestral scheduler and 75 inference steps, generating six consistent viewpoints per input. First, denoising and super-resolution are applied to improve sharpness and reduce noise, using ESRGAN to enhance spatial detail when needed. Next, background removal is performed using the Rembg toolbox to foreground the target object and produce cleaner object boundaries for subsequent processing. Background removal is performed using Rembg, and super resolution is applied using Real-ESRGAN (RealESRGAN_x4plus, 4× upscaling) to improve image quality prior to multiview

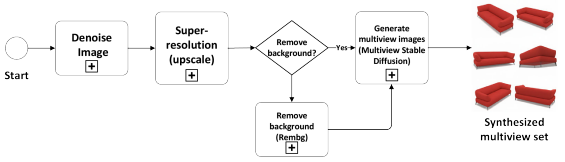

Figure 3: Phase 2: Image enhancement and multiview generation workflow.

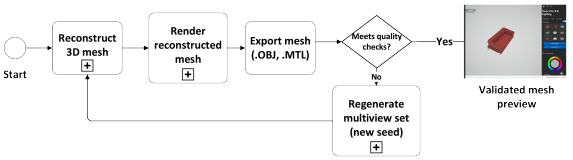

Figure 4: Phase 3: 2D-to-3D reconstruction workflow.

generation.

Following enhancement and segmentation, multiview synthesis is performed using Multiview Stable Diffusion to generate additional viewpoints of the same object. These synthetic views approximate photographs captured from different camera angles and provide the multi-perspective observations required for sparse-view 3D reconstruction. The overall sequence of enhancement, isolation, and multiview generation is summarized in Figure 3.

Phase 3: 2D-to-3D Reconstruction

Phase 3 converts the enhanced multiview image set into a 3D mesh representation. The processed views from Phase 2 are provided as input to a sparse-view reconstruction workflow (Instant Mesh), which generates a textured mesh by estimating object geometry from limited viewpoints. Reconstruction is performed using InstantMesh (instant_mesh_large.ckpt) with default configuration settings, operating on six synthesized views arranged in a 3×2 grid. Because reconstruction quality is sensitive to view consistency and segmentation quality, the resulting mesh is inspected using basic quality checks focused on geometric completeness and surface stability. When reconstruction artifacts are observed, the pipeline returns to Phase 2 to regenerate the multiview set with adjusted seeds, producing a new view collection for reconstruction. The reconstruction loop and decision point are illustrated in Figure 4. Final validated meshes are exported as `.OBJ` and `.MTL` files, preserving both geometry and associated material information for AR deployment.

The reconstruction process operates in a normalized coordinate space, and the metric scale is not inherently preserved. A configurable scale parameter is applied post reconstruction, allowing the generated mesh to be resized relative to the target scene.

Phase 4: 3D-to-AR Integration

Phase 4 prepares the reconstructed mesh for deployment in an AR environment. During AR integration, reconstructed meshes can be interactively rescaled to align with user defined spatial constraints and real-world placement requirements. The exported `.OBJ` and `.MTL` assets are con-

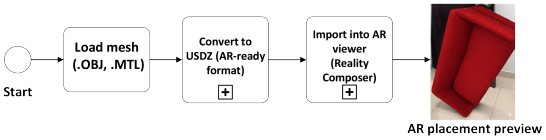

Figure 5: Phase 4: 3D-to-AR integration workflow.

verted into the USDZ format using Aspose 3D, enabling compatibility with common mobile AR workflows. After conversion, the asset is calibrated to ensure correct scale and orientation for placement in real indoor spaces. The calibrated USDZ model is then imported into Apple Reality Composer to support spatial anchoring and interactive visualization, including consistent placement within the physical environment and realistic rendering behavior. This integration sequence is summarized in Figure 5, and it produces an AR-ready asset that can be deployed on consumer-grade mobile devices for remodeling visualization.

This methodology establishes a practical baseline workflow for translating free-form user intent into AR-ready remodeling visualizations. Rather than proposing new foundation models, the contribution of this section is the structured integration of existing components into a reproducible sequence with clear intermediate outputs, enabling inspection and iteration before AR deployment. Evaluation procedures, parameter sensitivity observations, and quality outcomes are reported in the subsequent case study to assess visual plausibility, reconstruction robustness, and AR deployability in consumer-grade environments.

Unless otherwise specified, all models are used with default pretrained weights and configurations, with stochastic variation controlled through seed initialization.

Case Study: Results and Discussion

To evaluate the practical effectiveness of the proposed text-to-AR workflow, this study conducts a case study that examines how reliably visually coherent and dimensionally plausible 3D assets can be produced from limited two-dimensional inputs. The proposed methodology is fully executable, and the complete end-to-end pipeline was implemented and run as presented, from text prompting through AR-ready deployment. This evaluation is essential for home remodeling use cases where the objective is not fabrication-level precision, but the ability to translate homeowner intent into a shared visual artifact that supports clearer communication during early-stage design discussions. Accordingly, the case study assesses the pipeline's performance under varied input conditions, identifies key bottlenecks, and establishes baseline benchmarks that inform future improvements.

Experiment Overview

The primary objective of this experiment is to evaluate the robustness of the pipeline's 2D-to-3D reconstruction stage under diverse input conditions relevant to remodeling scenarios. Within the scope of this paper, the case study focuses on the most failure-prone and decision-critical portion of the workflow, Phases 2 and 3 (image preprocessing and multiview generation, and 2D-to-3D reconstruction), since these steps largely determine whether the final AR asset is usable for remodeling visualization. Sofa objects were selected because they are common in interior design tasks and present geometric challenges (curvature, occlusions, and complex silhouettes) that stress reconstruction quality in ways that are meaningful for AR-based space planning.

The dataset comprises sofa images organized into three categories namely a) Front-view images: frontal perspectives that preserve symmetry and provide clear visibility of primary features. b) Side-view images: lateral perspec-

Figure 6: Complete workflow visualization for four representative sofas. Each row shows: (a) original input image, (b) one representative view from the generated multiview set, and (c) final AR placement in a real room environment.

tives that introduce occlusion and asymmetry challenges; c) Synthetic images: controlled images collected for systematic evaluation under consistent conditions.

Real images were captured in front and side views to approximate practical acquisition conditions. In total, 12 distinct sofa models were evaluated to ensure variation in shape, style, and visual complexity. Synthetic sofa images were additionally collected from an online store to support controlled comparisons across consistent viewpoints and visual conditions.

To evaluate robustness under practical acquisition conditions, we applied controlled augmentations that vary image scale and quality. Each sofa image was resized to three resolutions (256, 512, and 1024 pixels) to assess sensitivity to input scale. For each resolution, two quality conditions were created. High-quality samples preserved sharpness and minimal noise to approximate favorable capture conditions. Low-quality samples were intentionally degraded using Gaussian blur (standard deviation = 2.0) with additional controlled noise to simulate common imperfections such as slight defocus and sensor noise, consistent with prior imaging literature (Roboflow, 2023; Gonzalez and Woods, 2023; Shorten and Khoshgoftaar, 2019). Controlled lighting variations were also introduced to reflect typical indoor capture variability. Figure 6 illustrates a representative sample processed through the complete workflow, showing the input image, preprocessing and background removal, multiview synthesis output, and the resulting reconstructed 3D mesh prepared for AR visualization.

All augmented inputs were processed through the full reconstruction workflow, including multiview synthesis, mesh reconstruction, and final rendering. This design enabled systematic comparison of pipeline performance across viewpoint type, resolution, and quality condition, providing an interpretable baseline for assessing reconstruction stability and identifying failure modes relevant to AR remodeling visualization.

Evaluation Metrics

To assess the proposed 2D-to-3D reconstruction workflow in a manner aligned with home remodeling visualization, we evaluate both (1) perceptual visual fidelity, which supports visually plausible AR previews for homeowner–professional alignment, and (2) geometric dimensional accuracy, which supports spatial plausibility during placement and scale interpretation in AR. Accordingly, we report one perceptual metric (SSIM) and two complementary dimension-based error metrics (NAE and NRMSE).

Perceptual visual fidelity (SSIM): The Structural Similarity Index Measure (SSIM), introduced by Wang et al. (Wang et al., 2004), evaluates image similarity using luminance, contrast, and structural components and is widely used in image synthesis and super-resolution research due to its correlation with human perception (Ledig et al., 2017; Zhang et al., 2018). In this study, SSIM is computed by comparing rendered projections of the reconstructed 3D mesh against the original input image and against its background-removed counterpart. The latter comparison reduces background influence and better isolates the reconstruction quality of the foreground object. SSIM values range from 0 to 1, where higher values indicate greater perceptual similarity.

Geometric dimensional accuracy (NAE and NRMSE): To evaluate dimensional plausibility, we adopt two metrics commonly used in object reconstruction and modeling studies (Tatarchenko et al., 2017; Wu et al., 2016). First, Normalized Absolute Error (NAE) measures the relative deviation between predicted and ground-truth dimensions (width, depth, and height):

$$E_{dim} = \left| \frac{D_{pred} - D_{true}}{D_{true}} \right| \tag{1}$$

where D_{pred} is the reconstructed dimension and D_{true} is the ground-truth measurement. This normalized form enables consistent comparison across objects of different sizes. Second, we use Normalized Root Mean Squared Error (NRMSE) as a scale-independent aggregate measure of dimensional accuracy:

$$NRMSE = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(D_{pred}^{(i)} - D_{true}^{(i)}\right)^2}}{\bar{D_{true}}} \tag{2}$$

where N denotes the number of evaluated samples and $\bar{D_{true}}$ denotes the mean of the ground-truth dimensions across the dataset. Lower NAE and NRMSE values indicate improved dimensional agreement.

Together, SSIM, NAE, and NRMSE provide a concise and complementary evaluation of visual plausibility and dimensional deviations. Results are reported across viewpoint type, resolution, and quality condition to characterize reconstruction robustness under varied input scenarios.

Results

Following data acquisition and augmentation, all samples were processed through the complete evaluation workflow. SSIM was computed to quantify perceptual fidelity by comparing rendered projections of the reconstructed meshes against the original input images and their background-removed counterparts. Geometric accuracy was assessed by comparing reconstructed object dimensions (width, depth, and height) against ground truth measurements across the 12 sofa models. Results are reported across image resolution (256, 512, and 1024 pixels), input type (front-view, side-view, and synthetic), and quality condition (high vs. low) to capture performance variability under realistic acquisition conditions. These results provide a baseline characterization of the tradeoff between perceptual plausibility and dimensional plausibility for AR-ready assets generated from limited inputs, which is critical for determining when such outputs are usable for early-stage remodeling communication.

Across all samples, the average SSIM between rendered meshes and original input images was 0.556, indicating

that the reconstructed outputs preserved key structural characteristics despite texture and mesh granularity limitations. When SSIM was computed against background-removed inputs, the average increased slightly to 0.564, suggesting that background complexity can depress perceptual similarity scores and that foreground isolation provides a clearer assessment of reconstruction quality. These SSIM values are consistent with ranges reported in prior single-input-view 3D reconstruction studies (Richardson et al., 2016; Wu et al., 2016). Dimensional accuracy was evaluated using NAE and NRMSE. Across the dataset, NAE values for width, depth, and height typically ranged from 20% to 60%, with width generally exhibiting greater stability. Table 2 reports the consolidated, scale-independent NRMSE values for each dimension. The results indicate that proportional structure is often preserved, while depth remains the most challenging dimension to recover reliably due to occlusion and limited visual cues.

Table 2: NRMSE by dimension across all evaluated sofa models.

Dimension	NRMSE
Width	30.98%
Depth	39.40%
Height	19.16%

Table 3: Per-sofa reconstruction quality across input conditions. SSIM is reported against original inputs. NAE is reported per dimension (W = width, D = depth, H = height).

Sofa	Input Type	Resolution (px)	SSIM (Orig.)	NAE_W (%)	NAE_D (%)	NAE_H (%)
S1	Front	512	0.325	30.0	70.0	2.6
S2	Side (Synthetic)	512	0.652	6.7	17.7	5.3
S3	Front	512	0.443	37.5	15.6	7.0
S4	Side	512	0.493	5.6	1.0	3.8
Average			**0.478**	**20.0**	**26.1**	**4.7**

To illustrate variability across representative cases and input conditions, Table 3 reports SSIM and dimensional errors for four sofa models at 512×512 resolution. The results show that height reconstruction is highly accurate (4.7% average error), while depth remains the most challenging dimension (26.1% average error), particularly for front-view inputs where occluded rear geometry must be inferred.

The results highlight the relationship between perceptual fidelity, geometric accuracy, and practical usability for AR remodeling visualization. Key observations are summarized below.

Visual plausibility is consistently maintained: The average SSIM values of 0.556 (against original inputs) and 0.564 (against background-removed inputs) indicate that reconstructed meshes preserve the primary structural features of the sofas. The slight improvement after background removal suggests that background complexity can suppress perceptual similarity scores and that foreground isolation provides a clearer assessment of reconstruction quality.

Depth remains the dominant source of dimensional error: Dimension-based evaluation shows larger discrepancies in depth compared to width and height. This outcome is expected when reconstruction begins from a single input view and relies on synthesized viewpoints, as occluded regions provide limited cues for recovering true depth.

Input view and resolution influence reconstruction stability: Reconstructions from front-view inputs are generally more stable than those from side views, where asymmetry and occlusion introduce additional ambiguity. Higher resolutions tend to improve SSIM and modestly reduce dimensional errors, although gains diminish at higher resolutions.

Implications for AR remodeling use: The pipeline produces visually coherent assets that support early-stage ideation, AR-based exploration, and communication of intent. However, the observed dimensional inaccuracies indicate that the current workflow is not intended for high-precision tasks such as fabrication or construction detailing.

Overall trade-off: The results demonstrate a clear trade-off between visual plausibility and geometric accuracy. The pipeline is effective for generating AR-ready previews for communication, while future improvements should prioritize scale awareness, depth reliability, and constraint-based reconstruction.

Conclusions and Future Research Directions

This paper implements an executable first-step text-to-AR workflow aimed at reducing the homeowner–professional communication gap in early-stage home remodeling by translating free-form user intent into an AR-ready 3D visualization. The primary contribution is the practical integration of established components (text-to-image generation, image preprocessing and multiview synthesis, sparse-view reconstruction, and USDZ-based AR deployment) into a reproducible pipeline with inspectable intermediate outputs that support iterative refinement and reduce reliance on manual 3D modeling expertise. While the full workflow is implemented end to end, the quantitative evaluation is intentionally focused on the most failure-prone and decision-critical stages that govern reconstruction quality and dimensional plausibility.

A sofa case study was selected as a conservative test object due to curved geometry and frequent occlusions, which stress multiview consistency and depth recovery more than common rectilinear interior elements. The results show that the implemented workflow can produce visually coherent assets suitable for AR ideation and communication, with average SSIM values of 0.556 against original inputs and 0.564 against background-removed inputs. However, geometric accuracy remains limited, particularly for depth, indicating that the current workflow is best positioned as an intent-alignment and early visualization tool rather than a substitute for precision measurement or fabrication-ready modeling. Overall, these findings support the feasibility of producing AR-ready previews from limited inputs and clarify the primary technical limitations that must be addressed in future work to improve dimensional plausibility and environment-aware placement.

Future work will focus on four directions: (1) environment-aware capture and spatial reasoning (e.g.,

leveraging mobile camera/LiDAR) to support scale- and fit-aware placement and basic clearance checking; (2) room-scale, multi-object scene generation with stronger geometric constraints and priors (standard dimensions and optional user measurements) to reduce depth ambiguity; (3) user-centered AR evaluation to quantify usability, perceived realism, and the impact on homeowner–professional decision-making; and (4) benchmarking direct text-to-3D methods within the same AR deployment constraints to compare dimensional plausibility and post-processing requirements against the staged workflow used in this study. Future work will also expand evaluation beyond sofas by collecting and labeling diverse furniture and interior asset datasets across multiple viewpoints to more rigorously assess generalizability and robustness. In addition, we are exploring user-captured imagery as input by integrating grounded detection and segmentation methods (e.g., Grounding DINO and Grounded SAM) to isolate real objects and evaluate projection fidelity in AR under practical conditions. Future work will also focus on incorporating explicit scale control into the pipeline, including user-defined dimensional inputs to guide scaling, automated dimension estimation from images, and scale-aware normalization prior to evaluation. Providing users with structured prompt templates that include dimensional cues is also expected to improve the generation of dimensionally plausible AR assets. Finally, we are investigating more controllable and open model alternatives to enable stronger user control over key design attributes such as dimensions and color.

References

Azuma, R. T. (1997). A survey of augmented reality. Presence: Teleoperators and Virtual Environments, 6(4):355–385.

Boopathy, K., Kalasapudi, V., Mantha, B. R., and Zou, Y. (2024). A survey-based investigation to understand the influence of an augmented reality-based application on the home remodeling market. In Construction Research Congress 2024, pages 426–435.

Brown, E. Q. (2020). How COVID-19 has changed home design. https://www.architecturaldigest.com/story/covid-19-home-design. Accessed: 2025-03-13.

Chen, Y., Liu, H., Zhang, X., and Lin, D. (2023). Diffusion models for text-driven interior design. ACM Transactions on Graphics (TOG), 42(2):1–15.

Coxworth, B. (2025). Augmented carpentry system brings furniture design into the real world. https://newatlas.com/vr/augmented-carpentry-system/. Accessed: 2025-04-06.

for Housing Studies of Harvard University, J. C. (2014). Emerging trends in the remodeling market. Technical report, Harvard University.

Fridman, A., Li, Y., Wang, C., and Funkhouser, T. (2023). Scenescape: Multi-view diffusion models for generating consistent 3d scenes from text. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 45(6):7689–7701.

Gkioxari, G. and Malik, J. (2019). Mesh R-CNN. ICCV.

Gonzalez, R. C. and Woods, R. E. (2023). Digital Image Processing. Pearson, 4th edition.

Ho, J. and Salimans, T. (2023). Diffusion models for 3d generation. Advances in Neural Information Processing Systems.

Höllein, L., Yang, Y., and Kopf, J. (2023). Text2room: Room-scale 3d reconstruction from text descriptions. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 5673–5682.

Höllein, L., Yang, Y., and Kopf, J. (2023). Text2room: Room-scale 3d reconstruction from text descriptions. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 5673–5682.

IKEA (2017). Ikea place. https://www.ikea.com/us/en/customer-service/mobile-apps/. Accessed: 2025-04-14.

Inc., H. (2018). Houzz introduces view in my room 3d. https://www.houzz.com/press/590/Houzz-Introduces-View-in-My-Room-3D. Accessed: 2025-04-01.

Joint Center for Housing Studies of Harvard University (2023). Improving America's housing 2023. https://www.jchs.harvard.edu/improving-americas-housing-2023. Accessed: 2025-03-16.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., and Shi, W. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4681–4690.

Lin, C.-H., Gao, J., Tang, L., Lee, Y.-C., Xu, J., Lai, W.-C., and Anandkumar, A. (2023). Magic3d: High-resolution text-to-3d content creation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 15234–15244.

Mildenhall, B., Srinivasan, P. P., and Tancik, M. (2020a). NeRF: Representing scenes as neural radiance fields for view synthesis. ECCV.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. (2020b). Nerf: Representing scenes as neural radiance fields for view synthesis. Proceedings of the European Conference on Computer Vision (ECCV), pages 405–421.

Poole, B., Jain, A., Barron, J. T., and Mildenhall, B. (2022). Dreamfusion: Text-to-3d using 2d diffusion. Advances in Neural Information Processing Systems (NeurIPS).

ReportLinker (2023). Home improvement services market - global outlook and forecast 2023-2027. `https://www.reportlinker.com/p06417077/`. Accessed: 2025-04-16.

Richardson, E., Sela, M., Or-El, R., and Kimmel, R. (2016). Learning detailed face reconstruction from a single image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 5553–5562.

Roboflow (2023). Data augmentation techniques for computer vision. Accessed: 2025-06-12.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 10684–10695.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., and Ho, J. (2022). Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems (NeurIPS).

Schönberger, J. L. and Frahm, J.-M. (2016). COLMAP: A general-purpose structure-from-motion pipeline. 3DV.

Seitz, S. M. and Curless, B. (2021). Multi-view stereo: A tutorial. Foundations and Trends in Computer Graphics and Vision.

Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6(1):60.

Tatarchenko, M., Dosovitskiy, A., and Brox, T. (2017). Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 2088–2096.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4):600–612.

Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., and Xiao, J. (2016). Learning 3d shape representations from 2d images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1919–1928.

Zhang, K. and Lee, Y. (2023a). Flexgen: Controllable multi-view image synthesis. CVPR.

Zhang, K. and Lee, Y. (2023b). Flexgen: Controllable multi-view image synthesis using diffusion models. Proceedings of the IEEE/CVF CVPR, pages 1254–1263.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. CVPR, pages 586–595.