Orchestrating Generative AI Paradigms With Human-in-the-Loop for 3D Generation

Emanuele Balloni, Lorenzo Stacchio, Marina Paolanti, Primo Zingaretti, *Life Senior Member, IEEE*, and Roberto Pierdicca

***Abstract*—Generative AI techniques are revolutionizing the creation of 3D and immersive content, yet challenges remain, such as achieving precise user control in 3D generation. Current text-to-3D and image-to-3D pipelines often produce outputs that deviate from user expectations, lacking the ability to refine or correct generated models effectively. To address these limitations, we propose Imagin3D, a novel human-in-the-loop (HITL) system that integrates Multimodal Large Language Models to enhance the controllability and adaptability of 3D content generation. Imagin3D leverages a Multi-View Question Answering module to evaluate the consistency of generated views with user-provided textual descriptions, enabling iterative refinement through guided inpainting while preserving multi-view consistency. This allows users to co-create 3D models, which are then synthesized into a final 3D asset using Neural Rendering. We validate Imagin3D through extensive quantitative evaluations and a comprehensive user study, demonstrating its effectiveness in improving usability, accuracy, and user satisfaction in interactive 3D generation tasks. Our results highlight the potential of HITL approaches to bridge the gap between AI-generated outputs and user intent, paving the way for more accessible and user-centered 3D generation workflows.**

***Index Terms*—Generative AI, image generation, 3D generation, human-in-the-loop, user intent, content customization.**

I. Introduction

GENERATIVE Artificial Intelligence (GenAI) algorithms are transforming the way we create 3D and immersive content. Recent advances in text-to-3D and image-to-3D generation enable AI to synthesize three-dimensional objects from high-level inputs, reshaping how virtual worlds, designs, and processes can be developed [1]. Despite its importance, creating geometrically accurate 3D models reflecting the user's specific requests, starting from a single image or text, remains a considerable challenge [2]. Despite state-of-the-art 3D generation and reconstruction methods that can produce high-quality results [1], [2], they often lack precise user 3D generation control, yielding outputs that do not align with what the user envisioned. Current approaches largely fall into two limited paradigms: (i) Text-to-3D pipelines that take a descriptive prompt but offer only coarse control over the result [2], and (ii) Image-to-3D reconstruction that builds a model from one or few reference images [3]. The former often struggles with fine-grained specifications (beyond the initial text prompt), while the latter limits the result to mimic the input image and can introduce artifacts when "lifting" 2D details into 3D geometry, potentially leading to unwanted results. In practice, this means users have little ability to correct or refine the generated model if it deviates from their intended outcome. Considering Image-to-3D methods, another critical aspect regards multi-view consistency: many Image-to-3D paradigms work by producing novel 2D images of the depicted object (novel views), that must be combined later into a 3D asset with Neural Rendering approaches [3]. Without specific techniques, a generative model that produces different views independently, usually a diffusion model [3], [4], often fails to keep them geometrically and texturally consistent. Moreover, making targeted edits to the generated content exacerbates this issue: a local change in one viewpoint (for instance, painting a door on the front view of a house) will not automatically reflect on the other views. Solutions to such a problem are non-trivial [2].

Given these limitations, there is a growing consensus that human-in-the-loop (HITL) approaches are needed to make generative 3D modeling more controllable and practical [2], [5]. Rather than leaving the generative process fully autonomous, a HITL framework keeps the user actively involved at key steps - guiding the AI model, correcting mistakes, and refining the output. Incorporating human feedback in this way can solve ambiguities and align the output with the user's intent [6], [7]. In this context, Multimodal Large Language Models (MLLMs) have emerged as a powerful facilitator for HITL creativity [8]. Modern MLLMs possess high multimodal understanding and therefore can act as intelligent assistants in the generative loop [9]. Recent works have proved how MLLMs can act as a "bridge" between users and complex content modalities like images and 3D assets, showing improvements in users' generative process [6], [7]. Despite these advances, existing approaches still offer limited support for explicit back-and-forth HITL design processes in which intermediate outputs are semantically analyzed, exposed to user inspection, and iteratively refined before final 3D reconstruction. In particular, prior pipelines for text-to-3D, image-to-3D, or interactive editing typically provide

Received 27 March 2025; revised 29 April 2026; accepted 16 May 2026. Date of publication 22 May 2026; date of current version 17 June 2026. Recommended for acceptance by V. Kim. *(Corresponding author: Emanuele Balloni.)*

Emanuele Balloni and Roberto Pierdicca are with the Department of Civil Engineering Building, Architecture, Università Politecnica delle Marche, 60131 Ancona, Italy (e-mail: e.balloni@pm.univpm.it).

Lorenzo Stacchio and Marina Paolanti are with the Department of Political Sciences, Communication, International Relations, University of Macerata, 62100 Macerata, Italy.

Primo Zingaretti is with the Department of Information Engineering, Università Politecnica delle Marche, 60131 Ancona, Italy.

This article has supplementary downloadable material available at https://doi.org/10.1109/TVCG.2026.3695993, provided by the authors.

Digital Object Identifier 10.1109/TVCG.2026.3695993

either direct generation with limited structured feedback, or refinement mechanisms without an explicit multi-view semantic validation stage.

To address these challenges, we present Imagin3D, a GenAI-driven system for 3D content creation with a HITL approach. Imagin3D pipelines different GenAI models with interactive user input and MLLM assistance, enabling users to iteratively design and co-create 3D models by interacting with the intermediate steps of the generative process, before the final 3D reconstruction. The system allows, starting from a single input image and textual prompt, to guide the user in the generation and refinement of multi-view 2D representations of the target 3D object, by integrating MLLMs, Image2Image, and Neural Rendering models. Within this workflow, the newly introduced Multi-View Question Answering (MVQA) module plays a central role by leveraging an MLLM to assess whether the generated views align with the user's specifications. By analyzing multiple perspectives of the object, MVQA applies a Davidsonian Scene Graph (DSG)-based approach [10] to reason about structural and compositional consistency (e.g., "Does the chair have four legs?"), thereby providing users with semantic feedback that can inform subsequent refinements. Exploiting the MVQA, Imagin3D enables MLLM-assisted refinement: users can request modifications, and the system translates these edits into visual structured inpainting instructions, ensuring multi-view coherence through our Score-Guided Inpainting (SGI) module. The refined 2D views are then synthesized into a complete 3D model using a neural rendering technique. The outcome is a collaborative workflow in which humans and GenAI work together to the creation of 3D assets, where users progressively steer the output toward their vision. We clarify that Human co-creation amounts to the interaction with the MLLM through the MVQA and the SGI stage, which allows users to both refine the prompts and generate multi-view images. The system has been implemented by integrating a Web Application, enabling users to perform each step of the system in a user-friendly manner. To validate the system, we conducted a comprehensive evaluation of the MVQA module and carried out an experimental user study involving 15 participants to assess the overall efficacy, efficiency, and usability of the system. The results indicate positive user adoption and demonstrate improved effectiveness of the generative models when guided through the proposed HITL workflow.

While recent works have explored text-to-3D, image-to-3D, and interactive editing pipelines, most of them either rely on direct generation with limited user steering or support refinement without an explicit multi-view semantic validation loop. In contrast, Imagin3D is designed as a HITL orchestration framework in which multi-view generation, MLLM-based multi-view question answering, guided inpainting, and 3D reconstruction are combined into a unified iterative workflow. Our goal is therefore not to introduce a new standalone 3D generator, but to improve controllability and alignment to user intent through structured feedback and iterative user intervention.

Our contributions can be summarized as follows: (i) We introduced Imagin3D, a novel system that orchestrates GenAI paradigms, exploiting an innovative HITL framework to enhance human-AI collaboration for 3D generation; (ii) We defined a novel application of MLLMs for MVQA, alongside a Consistency-Aware Inpainting Mechanism, to guide the user in the 3D generation process; (iii) We developed a UI for the system that lets end users perform all the steps in an easy-to-use manner; (iv) We conducted a comprehensive quantitative evaluation of the novel MVQA module on the Google Scanned Objects (GSO) observing the effectiveness of our approach; We carried out a user study assessing Imagin3D's effectiveness in guiding the user during the 3D generation process, with a Human-Computer Interaction perspective.

II. Related Works

In this Section, we review the state-of-the-art methods falling nearest to out area of contribution according to 3D Generation, MLLM assistance and HITL AI systems.

a) Text-to-3D and Image-to-3D Generation: Early approaches to text-conditioned 3D modeling attempted to directly learn mappings from text to shape using paired datasets, but progress was limited by the scarcity of large-scale text-3D data [11]. Recent advances have shifted toward leveraging powerful 2D generative models and neural implicit representations to bridge this gap [3], [4]. One prominent line of work leverages Neural Rendering methods, such as Neural Radiance Fields (NeRFs) and SDF-based approaches, optimizing them so that renderings match a target description [3], [4]. In conjunction with text-to-3D, image-to-3D generation has emerged as another key area of research. Works such as [3], which employs a diffusion-based mechanism to generate 3D content from a single image. Usually, these approaches fine-tune a 2D diffusion model to generate novel views of an input image, which serve as a 3D prior for higher-quality view reconstructions. While existing methods demonstrate considerable capabilities, they struggle with consistency across views and following textual prompts in varied scenarios. The proposed approach improves 3D synthesis by leveraging diffusion priors and incorporating human guidance and editable outputs, addressing both fidelity and editability through generative AI techniques.

b) MLLM assisted Generation: MLLMs are increasingly used as coordinators in multimodal generation, guiding workflows and refining outputs beyond text and could also serve as critics, iteratively refining generative outputs by detecting inconsistencies or missing details. On this line, [12] introduced iterative LLM-guided refinements for text-to-image generation. This kind of two-phase generate-and-refine loop has been shown to significantly improve adherence to the prompt, and the same principle can be applied to multi-view generation. Also, MLLMs can be applied to various forms of multimodal assistance, such as Visual Question Answering (VQA). VQA methods, however, are often limited to single images, may struggle with questions requiring multi-view or 3D understanding. Multi-view VQA addresses this by supplying multiple images to the model, necessitating MLLMs capable of processing such inputs. This was investigated in [13], where a multi-view MLLM was fine-tuned on tasks like 3DMV-VQA, showing strong capabilities in spatial reasoning and complex visual understanding. Building on this line of work, *Imagin3D* utilizes MLLMs to support MVQA on

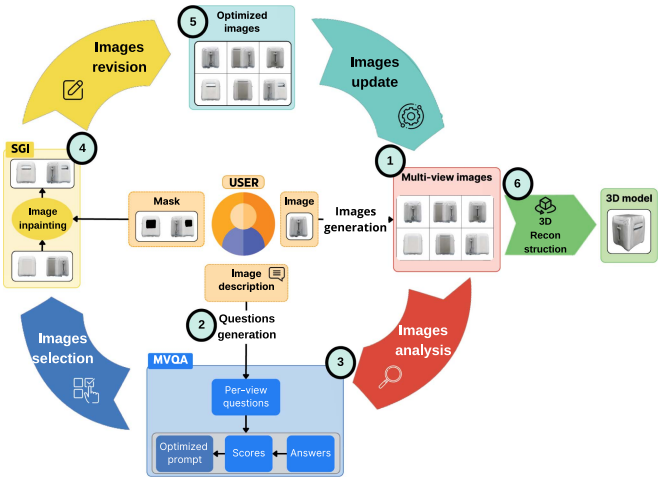


Fig. 1. Overview of the Imagin3D system.

generated views and 3D models, adopting a novel approach compared to existing literature. Imagin3D automatically generates questions and answers based on the rendered images to detect missing details relative to the user's initial textual description. To ensure semantic coherence, it employs the DSG score. Identifying these missing elements aids users in performing Generative Manipulations on the views.

c) Human-in-the-Loop AI Systems for 3D Modeling: Integrating human feedback into the generative process can substantially improve the quality and usability of AI-created content [2], [14]. Prior interactive design tools in both 2D and 3D domains have shown that a HITL approach helps guide generative models toward the user's intent. An example amounts to DragGAN [14], which allows users to interactively drag points on a GAN-generated image to deform objects, easily steering the generation. In the 3D domain, authors of [2] employs a two-stage generative process, allowing the user to intervene between stages, making adjustments before the final refinement (e.g, textual semantic editing and part combination). Considering modifications made on 2D views of a 3D scene, approaches like Instruct-NeRF2NeRF allow the user to provide textual instruction (e.g., "make the chair legs wooden") [15]. Imagin3D was built on top of such an idea, extending the generative paradigm by incorporating a HITL feedback mechanism. Instead of relying on one-shot generation, it enables iterative refinement where users can identify and correct issues, such as distortions or misinterpretations, using GenAI techniques. A feature-level comparison with prior works is reported in the Appendix (Section B-A), available online, to clarify how Imagin3D differs in terms of HITL interventions and implemented components.

III. Methods

Imagin3D is a system designed for user-guided 3D model generation from a single image and prompt. By incorporating a HITL and GenAI approaches, it aims to enhance the accuracy and consistency of 3D generation. The system integrates multiple GenAI components that assist users throughout the entire generation process. Its architecture, illustrated in Fig. 1, comprises three core modules: MVQA, SGI, and 3D Reconstruction (3DRec).

Novel views are generated with a Multi-View Diffusion (MVD) model. The MVQA module evaluates user prompt adherence with the generated images through a scoring mechanism. The SGI module allows users to review the MVQA output, including scores, questions, and answers, while providing an optimized prompt to facilitate inpainting. Additionally, SGI enables users to refine images to better align with the initial request, preserving multi-view consistency. Finally, the 3DRec module reconstructs the 3D object using the refined images, by leveraging a Neural Rendering approach. These modules are pipelined to achieve multi-view consistency, fast refinement, and prompt fidelity in the 3D reconstruction process. Imagin3D is implemented as a web-based application featuring an intuitive user interface that enables seamless visualization, interaction, and processing. The system UI consists of three main web pages, each corresponding to one of the core components. The application has been developed in Python and deployed on a web server using Gradio.[1]

To validate the system, we conducted an evaluation of the MVQA scoring procedure to assess the effectiveness of the proposed approach. Additionally, a comprehensive user study was performed to evaluate the system's usability, efficacy, and efficiency. The results of these evaluations are presented in Sections IV and V. The adopted validation strategy was intentionally twofold. On one end, we evaluated the MVQA component, implemented as a novel module of the Imagin3D system. On the other end, the remaining stages (multi-view diffusion, guided inpainting, and neural 3D reconstruction) rely on SOTA open-source methodologies. Specifically, we integrated: (i) a Multi-View Diffusion model [3], responsible for generating six novel views per object; (ii) a grid-based latent diffusion inpainting framework adapted from [16]; and (iii) a neural surface reconstruction approach based on NeuS [17]. Each of these modules was adopted as a robust building block, allowing the study to focus on validating the HITL integration rather than re-assessing each method's baseline performances [18]. We designed Imagin3D as a human-AI co-creation system to maintain methodological validity, rather than as a collection of independent GenAI approaches. This aligns with recent human-centered evaluations in GenAI systems, where user studies replace isolated benchmarks to assess usability and perceived fidelity under realistic interaction conditions [19], [20], [21].

A. Multi-View Question Answering (MVQA)

The MVQA module serves as the basis for multi-view consistency evaluation in Imagin3D. The system starts with a single user-provided image and employs a MVD model to generate six novel views of the object [3]. Given that a single image is typically insufficient for a complete 3D description, since details such as side or back features may not be visible, users can

[1]Gradio framework: https://www.gradio.app/

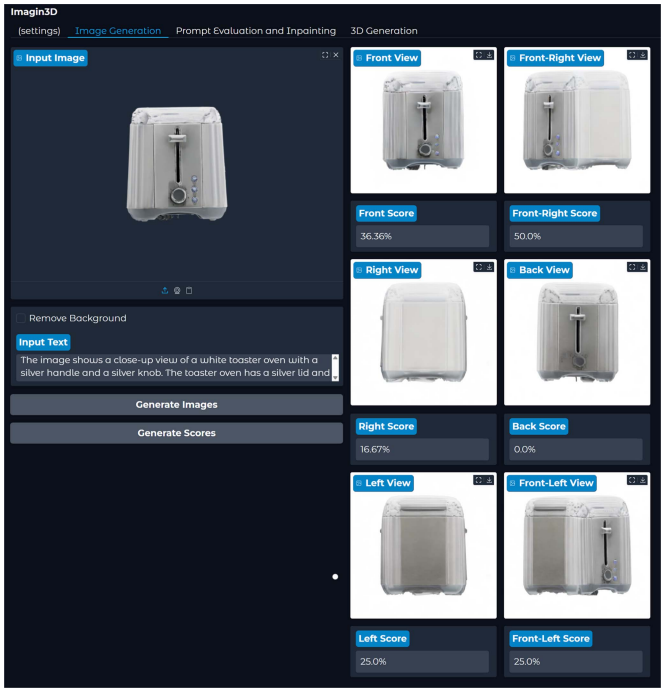


Fig. 2. "Image Generation" tab of the Imagin3D system. Given a single image and a textual description, 6 novel views are generated, and a text-image matching score is given to each image, letting users assess the text-image adherence.

supplement the input with a textual description, detailing the object from multiple viewpoints.

The primary goal of MVQA is to assess the consistency of the generated views with the provided textual description, helping users in the subsequent image refinements through GenAI methods. This is achieved through a series of structured steps, detailed in the following sections. The MVQA module is integrated into the UI of the Imagin3D system via the "Image Generation" tab. Fig. 2 shows an example of the interface.

a) Per-View Prompt Generation (PVPG): Starting from the user's input prompt z, six viewpoint-specific prompts z_c are generated using a LLM (LLama 3.3-70B in our case), each describing the object from a different perspective: $c \in \{\text{front, front-right, right, back, left, front-left}\}$. This ensures that each view is described in a localized and detailed manner. The viewpoint-specific prompt extraction is formalized as:

$$z_c = f_{\text{view}}(z, c, LLM), \quad c \in [1, 2, \ldots len(c)] \tag{1}$$

where f_{view} is the function that refines z to generate z_c, the prompt tailored for viewpoint c according to a MLLM. The function f_{view} instructs the LLM to extract viewpoint-specific descriptions from the comprehensive prompt. Its objective is to produce individual captions for each perspective-namely "front", "front-right", "right", "back", "left", and "front-left". These descriptions focus solely on the visible characteristics of the object from each respective angle, ensuring that details not observable from a given perspective are excluded. To achieve this, f_{view} begins by generating a prompt for the LLM that instructs it to create isolated descriptions for each viewpoint based on the comprehensive description z. The prompt specifies

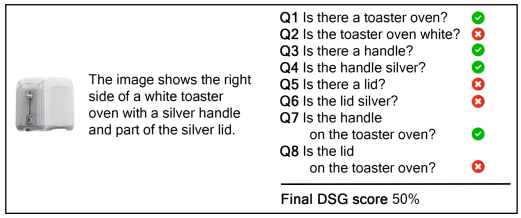


Fig. 3. Example of questions, answers, and DSG scoring based on the front-right description and generated image.

the need for exclusive descriptions by view, explicitly instructing the model to omit any details not visible from the specified angle. The model is then guided, through a meta-prompt, to output each viewpoint's description in a standardized format: $<view>:<caption>$ for each of the requested view. Furthermore, by instructing the model to avoid cross-referencing other views, f_{view} generates precise, isolated descriptions that focus on individual perspectives. For instance, the "front-right" view will only describe features observable from that angle, without inferring or referencing the "right" or "back" views. In this manner, f_{view} constructs a consistent, modular set of descriptions that collectively provide a complete understanding of the object, essential for applications requiring accurate multi-view analysis or evaluation (full implementation details can be found in the Appendix B-B, available online).

b) Questions Generation: In our system, a critical component involves feeding visual feedback to the LLM to evaluate and improve prompt-image consistency. This visual feedback is quantified through a consistency score that evaluates how well the generated images are consistent with the prompt descriptions. We also argue that the consistency score should be detailed enough for the LLM to infer ways to refine the candidate prompts. In practice, we defined an evaluation system through the DSG score, which assesses prompt-image consistency using a question generation and answering approach inspired by TIFA [22]. Specifically, DSG decomposes the view-specific prompts into atomic, unique binary questions that are organized into semantic dependency graphs. This process is repeated for each of the considered views. We use only the textual prompt as input for question generation, excluding the image. While the prompt expresses the user's intended target outcome, including the image would over-constrain the process to specific visual details of the input. Relying solely on text ensures that evaluation questions reflect the desired output rather than the content of the input image. To illustrate how our question generation and scoring mechanism works, consider the example in Fig. 3. From the view-specific prompt, a set of DSG binary questions are generated, such as: (i) Is there a toaster oven? (ii) Is the toaster oven white? (iii) Is there a handle? (iv) Is the handle silver? In this example, questions (ii) is dependent on question (i) and question (iv) is dependent on question (iii); thus, (i) serves as a prerequisite to validate (ii) and (iii) serves as a prerequisite to validate (iv).

c) Visual Question Answering (VQA): Once questions are generated for each view, they are paired with the corresponding MVD-generated images and processed by an MLLM (LLama

3.2 Vision-11B in our case). This model acts as a VQA system, providing, for each view, responses that determine whether the generated images align with the descriptions. This refinement mechanism is to address inconsistencies and inadequacies in the generated views.

Correct answers receive a score of 1, while incorrect or missing details score 0. In the example, the VQA model correctly identifies the presence of the toaster, the handle, its color and position on the toaster, attributing to every correct answer a score of 1. On the other hand, it also acknowledges the incorrect color of the toaster, and the absence of a silver lid, giving it a score of 0. The overall DSG score for the prompt-image pair is computed by averaging the individual VQA scores across all questions (and then presenting it as a percentage). In this case, the DSG score is 50%, reflecting the degree of alignment between the prompt and the visual content, providing a more detailed, question-based metric for evaluating prompt-image consistency.

For flagged views (i.e., those with a score $s_c < 100\%$), a prompt optimization step is performed to improve fidelity and alignment with the original input specifications. This is achieved by constructing a meta-prompt z_{meta} that consolidates evaluation metrics and feedback to guide the refinement process. The meta-prompt z_{meta} is formulated as:

$$z_{\text{meta}} = f_{\text{meta}}(q_c, a_c, s_c), \tag{2}$$

where q_c represents the set of descriptive questions associated with the viewpoint-specific prompt z_c, a_c is the set of answers derived from the generated image x_c^{gen} during the evaluation phase and s_c represents the view consistency score for viewpoint c. f_{meta} constructs a detailed textual representation summarizing discrepancies between the expected attributes (derived from q_c) and the observed attributes (represented by a_c). The meta-prompt is detailed in the Appendix.

Using the meta-prompt z_{meta}, our considered MLLM generates an updated viewpoint-specific prompt z_c^{new} that aims to address the identified shortcomings and help the user in the subsequent steps. This is formalized as:

$$z_c^{\text{new}} = f_{\text{refine}}(z_c, z_{\text{meta}}, MLLM), \tag{3}$$

where f_{refine} integrates the feedback encapsulated in z_{meta} into the original prompt z_c. The refinement process ensures that the updated prompt z_c^{new} emphasizes attributes or geometric details that were previously misrepresented or omitted in the generated image x_c^{gen}. This process is repeated for each view, providing novel prompts that are more accurate to the initial user description and that can aid the user in modifying the image through the SGI module. A Visual example which clarifies this entire process is reported in Appendix B-B, available online. A thorough evaluation of this approach has been performed through an experimental setting, detailed in Section IV.

B. Score-Guided Inpainting (SGI)

After the MVQA phase, images can be selected for further analysis and enhancement to better align with their corresponding descriptions before generating the 3D model. This refinement is achieved through the SGI module, which enables

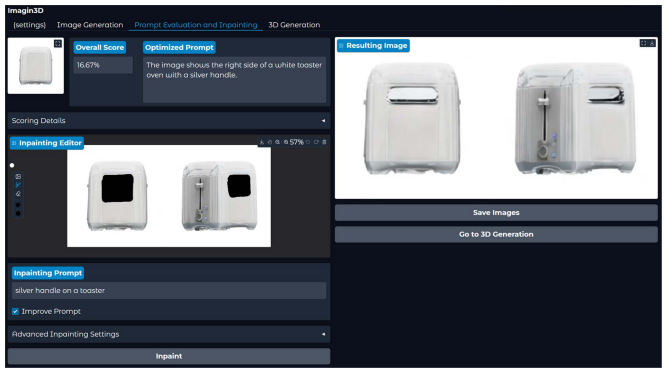


Fig. 4. Overview of tab "Prompt Evaluation and Inpainting".

image optimization with a HITL approach. The SGI module is integrated into the Imagin3D UI via the "Prompt Evaluation and Inpainting" tab (Fig. 4).

The module initially presents users with an overview of the scoring results for a selected view, including the detailed questions and answers generated by the MVQA module (this can be found under the "Scoring Details" tab, which opens with a simple user click; its visualization is reported in Appendix B-A, available online). Additionally, the optimized per-view description prompt is displayed to guide the user during inpainting. These elements are positioned in the top-left section of the interface (Fig. 4). After reviewing the image-prompt adherence based on the provided questions, answers, and scores, users can proceed to the inpainting phase (bottom-left of the interface). In this phase, users specify the desired modifications via a textual prompt and define a mask over the image regions to be altered using an integrated image editor (In our experimental setting, the masks were hand-drawn, but we integrated SAM to support users in automatic masking [23]. More details can be found in Appendix B-A, available online). Optionally, to support users who are not aware or experts of prompt engineering principles, an automatic prompt augmentation mechanism was implemented, to expand simple prompts into more detailed and accurate ones, following the same augmentation principles as in [24] (full details in Appendix B-C, available online).

To ensure multi-view consistency across different image viewpoints, we adopt a grid prior approach inspired by [16]. This method involves tiling multiple images into a 2×2 grid and processing them through a diffusion-based inpainting model (in our case, Stable Diffusion [25]), as a single image. Formally, let $I_{i,j}$ represent the image at position (i, j) in the grid, where $i, j \in \{1, 2\}$. The combined grid image G is defined as:

$$G = \begin{bmatrix} I_{1,1} & I_{1,2} \\ I_{2,1} & I_{2,2} \end{bmatrix} \tag{4}$$

Correspondingly, let $M_{i,j}$ denote the mask for $I_{i,j}$, and $P_{i,j}$ the prompt describing the desired inpainting for $I_{i,j}$. The inpainting model D then processes the grid as:

$$\hat{G} = D(G, M, P) \tag{5}$$

(a) Multi-view Consistent Inpainting. A 2×2 grid of images is masked in the same object region across different views, resulting in a consistent output adherent to the prompt.

(b) Visual Adaptation Inpainting. A single masked view in the 2×2 grid yields an output coherent with unmasked views while applying modifications based on the prompt.

Fig. 5. Examples of consistency-aware inpainting using *"purple wooden sofa frame"* as prompt.

where M and P are the aggregated masks and prompts for the grid, respectively. By aggregated masks M we refer to a mask composed of all the individual view masks within the considered 2×2 image grid, as drawn by the user (the dark masked areas in Fig. 4). The output $\hat{G}$ is the inpainted grid, from which individual images $\hat{I}_{i,j}$ are extracted. This approach ensures that inpainting is consistent across multiple views, as it enables the model to consider the spatial and contextual relationships between the tiled images during the inpainting process, based on the inpainting mask. In particular, two inpainting mechanisms have been implemented: Multi-View Consistent Inpainting and Visual Adaptation Inpainting. The former, inspired by [16], involves creating a mask across the 2×2 grid that targets the same object region in each image. Given a user-defined prompt p, the model inpaints all views simultaneously, ensuring coherence among different perspectives:

$$\hat{I} = D(I, M, p), \tag{6}$$

where I represents the input image grid, M denotes the mask, and $D(\cdot)$ is the diffusion-based inpainting function. The latter Visual Adaptation Inpainting, here introduced, follows a different strategy, by selectively masking a subset of the views within the 2×2 grid. Given a prompt p, the model applies modifications only to the masked areas while preserving the visual coherence with the unmasked views:

$$\hat{I}_m = D(I_m, M_m, p), \tag{7}$$

where I_m and M_m refer to the masked images and masks, respectively. This approach is particularly advantageous in ensuring consistency even when different views require distinct modifications.

Examples of both inpainting strategies are presented in Fig. 5. It is worth noting that Visual Adaptation Inpainting is a novel application of grid-based inpainting, which allows users to correct from a few views while retaining coherence for non-masked ones, instead of requiring multi-view masking from all viewpoints, making an addition towards more usable and efficient HITL pipelines. This process remains entirely transparent to the user, but is critical for maintaining multi-view consistency throughout the inpainting process. After creating the mask, the SGI module presents users with different inpainting options, allowing to fine-tune hyperparameters, such as guidance scale, number of diffusion steps, and inpainting strength. These settings are targeted towards more advanced users, seeking further customization. After initiating the inpainting process, the modified output image appears on the right side of the interface, allowing users to evaluate the result. If satisfied, they can save the updated images for 3D reconstruction; otherwise, they may adjust the inpainting settings, prompt description, or mask and perform the inpainting process again. The SGI module supports image refinement across all view-dependent images, aiming to maximize adherence to the user's description.

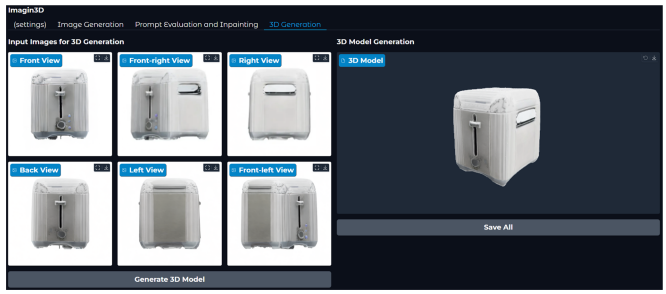


Fig. 6. "3D Generation" tab interface of the Imagin3D system.

C. 3D Reconstruction (3DRec)

Once users finalize the inpainting adjustments, the six refined views are fed into a 3D reconstruction model (NeuS in our case [17]), which generates a high-fidelity 3D representation of the object. The 3DRec phase is integrated into the Imagin3D system through the "3D Generation" tab (Fig. 6). From this interface, users can visualize, download, and interact with the generated 3D model, providing comprehensive ways to assess reconstruction quality.

IV. Multi-View Question Answering Experiments

A. Experimental Protocol

a) Evaluation Dataset: We used the GSO dataset [26] to perform the experiments. GSO provides a high-quality collection of 3D-scanned household items, and is widely used as a benchmark for single-image to 3D reconstruction tasks [27], [28], [29]. The dataset comprises diverse, photorealistic scans of everyday objects, offering realistic textures, accurate geometry, and a variety of poses and lighting conditions. The GSO dataset serves three main purposes: (1) it provides 16 ground truth (GT) images of each object from different angles, (2) it provides a single input reference image, which is used to guide the generation of novel views, and (3) it enables accurate testing thanks to the high-quality scenes provided. Following the setup in [29], we randomly select 30 scenes from GSO. For each object, a

single input image of resolution 256× 256 is rendered, serving as the basis for generating multi-view images and evaluating the consistency of our reconstructions.

b) Metrics: We evaluate Imagin3D using a set of established metrics across Natural Language Processing and Multimodal Computer Vision. BLEU measures n-gram overlap between generated and reference prompts, capturing lexical similarity [30]. METEOR assesses both precision and recall with synonym and stemming capabilities [31]. ROUGE computes n-gram overlap focusing on recall [32], and SPICE evaluates semantic content based on scene graphs, assessing object relationships [33]. CLIPScore quantifies visual-text coherence [34], while RefCLIPScore compares generated views with a reference 3D object view. The DSG Score is used for VQA and evaluates how well generated views capture scene elements when answering questions about object relationships.

c) Implementation Details: Our methodology relies on three core models with tailored configurations. For initial prompt generation, we use LongLLaVA-9B [35], optimized for multimodal causal language modeling, with 4096 hidden units, 32 attention heads, and a `silu` activation function. For per-view prompt generation and VQA, we employ Llama3.3-70B [36] and Llama 3.2-Vision, respectively. The Era3D model is used for the MVD component, as it is state-of-the-art in single-image-to-3D reconstruction. We adopt recommended hyperparameters for Era3D to ensure optimal performance. Regarding the 3D reconstruction, following [3], [37], we use a two-stage process. First, we generate initial 3D reconstructions using NeuS [17], then we apply a texture refinement step to enhance surface realism. All experiments were conducted on an NVIDIA A6000 GPU with 48 GB VRAM and 512 GB of RAM on a Ubuntu 22.04 system.

B. Results

In this Section, we present both quantitative and qualitative results obtained in our experimental setting of the MVQA module. To determine the most suitable MLLM for MVQA (in our case, LLAMA3.2-Vision), we conducted a preliminary evaluation on four different MLLMs. Results can be found in Appendix C, available online, together with further analysis on per-view values and more qualitative and quantitative results.

1) Quantitative Analysis: To create a baseline for the evaluation of the optimized prompt in out MVQA module, we generated a GT caption for each GSO object, by taking the GT set of n images SI $= \{x_1, x_2, \ldots, x_{16}\}$ and generating a textual description z_{long} that encompasses all n input images. This is formalized as following.

$$z_{\text{long}} = f_{\text{prompt}}(\text{SI}, MLLM), \tag{8}$$

where f_{prompt} is implemented using a MLLM [35]. Prompt details are available in the Appendix. The function f_{prompt} extracts semantic features from the input images and aligns them into a consistent textual representation. The z_{long} text is then used as the starting description for the MVQA process, together with the front view image extracted from the GT set of images. Starting from the z_{long} description, we perform the PVPG phase of the MVQA module to generate 6 view-specific GT descriptions, starting from the initial single GT description. The metrics extracted from 6 per-view descriptions are used as a baseline for our comparison. The same process is applied to the 6 images generated by the MVD model only, generating a single textual description of the images at first, and then creating 6 view-dependent descriptions from said text. These are compared against the optimized prompts generated at the end of the MVQA module, using the baseline prompts as GT.

TABLE I
COMPARISON OF PROMPT STRATEGIES FOR THE MVQA MODULE. BASELINE PROMPT: PER-VIEW PROMPT EXTRACTED FROM GT CAPTION. MVD PROMPT: PROMPT EXTRACTED FROM THE IMAGES GENERATED BY THE MVD MODEL. OPTIMIZED PROMPT: PER-VIEW PROMPT GENERATED WITH IMAGIN3D. LAST COLUMN SHOWS DIFFERENCES BETWEEN MVD AND OPTIMIZED PROMPT VALUES.

Metric	Baseline Prompt	MVD Prompt	Optimized Prompt (Ours)	Gain
BLEU-1	13.07%	**12.86%**	10.06%	–2.80%
BLEU-2	9.21%	**7.15%**	6.71%	–0.44%
BLEU-3	6.75%	4.06%	**4.53%**	+0.47%
BLEU-4	5.15%	2.36%	**3.06%**	+0.70%
METEOR	11.74%	9.31%	**10.61%**	+1.30%
ROUGE	22.96%	17.88%	**20.45%**	+2.57%
SPICE	23.78%	12.39%	**20.39%**	+7.99%
CLIPScore	73.66%	73.11%	**73.85%**	+0.74%
RefCLIPScore	76.36%	71.76%	**76.85%**	+5.09%

As shown in Table I, the Optimized Prompt strategy outperforms the MVD Prompt in the majority of evaluation metrics, underscoring the effectiveness of our approach. For the BLEU scores, which assess n-gram precision, the Optimized Prompt demonstrates improvements specifically for BLEU-3 and BLEU-4, though BLEU-1 and BLEU-2 show slight decreases. BLEU-1 and BLEU-2, however, are generally considered the least accurate of the BLEU scores in evaluating complex scene descriptions, as they primarily capture basic lexical overlap and are less indicative of semantic accuracy. The Optimized Prompt achieves higher performance across METEOR, ROUGE, and SPICE metrics, indicating better recall and F1-scores, particularly with a SPICE score of 20.39% which reflects enhanced semantic understanding and object relationship encoding. For visual-textual coherence, the Optimized Prompt outperforms in both CLIPScore and RefCLIPScore, with values reaching 73.85% and 76.85%, respectively, compared to lower scores in the Baseline and MVD prompts.

These gains demonstrate improved alignment between the generated prompts and reference views, essential for producing high-quality 3D reconstructions. It is interesting to note that our methods outperform even the baseline in visual-textual coherence. This outlines how the prompt optimization is able to retain high semantic similarity about the images while maintaining textual understanding. These results underscore Imagin3D consistent ability to deliver more accurate, semantically rich, and visually coherent prompts for single-image to 3D generation, capturing intricate scene elements effectively while maintaining strong alignment with reference views.

2) Qualitative Analysis: Fig. 7 presents some qualitative results obtained from the *toaster* scene in the GSO dataset (additional examples are included in Appendix C-C, available online). In these examples, we compare the view-dependent

View	Method	Image	Description
Front	MVD		A sleek, silver coffee maker with a black top and a shiny finish. The top features a lid that can be opened and closed.
	Ours		A white toaster oven with a silver knob and a silver handle on the front, the silver lid is also visible.
Front-right	MVD		The right side of the silver coffee maker is visible, showcasing its shiny finish. The edge of the black top is also visible, as is a small portion of the clear water reservoir window on the bottom.
	Ours		The right side of a white toaster oven with a silver handle and part of the silver lid.
Right	MVD		A view of the right side of the coffee maker, displaying its silver finish and the clear water reservoir window on the bottom.
	Ours		The right side of a white toaster oven with a silver handle.
Back	MVD		The back of the coffee maker appears as a plain silver surface with no visible buttons or controls.
	Ours		The back of a white toaster oven.
Left	MVD		A view of the left side of the coffee maker, displaying its silver finish and the clear water reservoir window on the bottom.
	Ours		The left side of a white toaster oven.
Front-left	MVD		The left side of the silver coffee maker is slightly visible, showcasing its shiny finish. The edge of the black top is also visible, as is a small portion of the clear water reservoir window on the bottom.
	Ours		The left side of the front of a white toaster oven with part of the silver knob and part of the silver lid.

Fig. 7. Qualitative results on the *toaster* scene. Prompts extracted from the MVD reconstruction and from Imagin3D prompt optimization are shown.

descriptions generated from the MVD images and the optimized description generated by the MVQA module, together with the original MVD-generated images and the images optimized through inpainting.

In particular, the optimized prompts generated by Imagin3D demonstrate a high degree of accuracy and detail across multiple viewpoints, resulting in a cohesive and consistent multi-view reconstruction. Starting from the following initial description: *"The image shows a close-up view of a white toaster oven with a silver handle and a silver knob. The toaster oven has a silver lid and a silver handle on the side"*, in the front view, our optimized description correctly identifies the white toaster oven with a silver knob and handle, aligning perfectly with initial description, while the MVD prompt incorrectly describes a silver coffee maker. Similar differences can be found in the other views, such as the front-right view, where the optimized prompt accurately describes the silver handle and part of the lid, while the MVD prompt deviates by describing a coffee maker's shiny finish and water reservoir. The qualitative results, paired with the improvement in metrics, reflect Imagin3D capability to create prompts that support a clearer and more accurate multi-view understanding of the 3D objects, aiding users in the enhancement process.

V. Human Evaluation Experiments

In this Section, we detail the experimental phase performed to investigate whether the proposed Imagin3D system could effectively be used for 3D generation through Generative AI assistance. This study was done considering the need to quantify how usable a GenAI system is from a human-centric perspective [2]. In particular, we describe the adopted Apparatus, the Experimental Process, the Participants, and finally, the Measurements adopted to evaluate the different variables under analysis. Additional experiments are reported in Appendix C-D, available online.

A. Participants

We enrolled 15 participants for our study. Considering our within-subject design with four Experimental Scenarios, each participant equally participated in all of them [38]. This sample size aligns with prior research, which has demonstrated that a similar number of participants is sufficient to identify over 80% of relevant interface properties and design issues [39]. Each participant completed a pre-experience questionnaire covering demographic details and their familiarity with 3D modeling and Generative AI. Our sample exhibited a mean age of 29 years (SD = 3.59, range = 25-38) and was primarily from Italy, with one participant from India. The educational background of the sample was distributed as follows: 53.33% (n = 8) held a Master's degree, 33.33% (n = 5) a Doctoral degree, and 13.33% (n = 2) a Bachelor's degree. Understanding 3D objects confidence was average, with a mean score of 3.33 (SD = 1.40), and slightly more than half of the participants (53.33%) had prior experience using 3D software editing tools such as MeshLab, Metashape, or Blender. However, proficiency in these tools remained low, with a mean score of 2.27 (SD = 1.28). A smaller proportion (33.33%) had experience searching for ready-to-use 3D models in databases like Sketchfab, and those who had used such databases reported moderate satisfaction with the customization options (M = 3.00, SD = 0.63). Generative AI familiarity was generally high, with all participants reporting awareness of the technology and 86.67% (n = 13) having used it. A similar phenomenon was found for Large Language Models, which was reported by 86.67% (n = 13) of participants, while familiarity with Large Reconstruction (3D reconstruction) models was slightly lower (46.67%). Knowledge of text-to-image models was also substantial (66.67%, n = 10). Among those who had used Generative AI tools, self-rated proficiency was moderate to high, with a mean score of 3.75 (SD = 0.87). These results indicate that our participants have a varying degrees of practical skills. This allow us to measure how differences in input quality, capturing robustness under realistic user conditions.

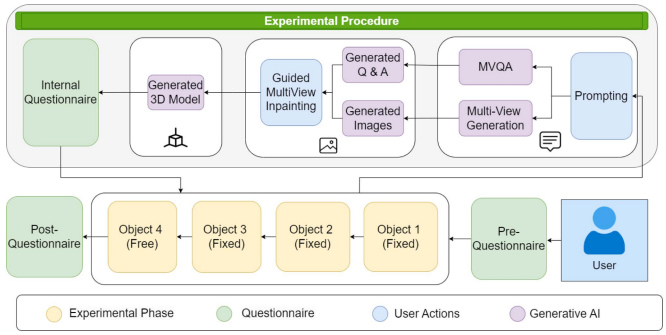


Fig. 8. Experimental Process.

B. Experimental Process

The adopted experimental procedure, illustrated in Fig. 8 and inspired by prior methodologies [19], [20], [40], was designed to capture both standardized and open-ended user interactions with Imagin3D. Each session began with informed consent, a demographic survey, and a pre-questionnaire assessing participants' background knowledge and familiarity with 3D modeling and Generative AI (more details in Appendix A-E, available online).

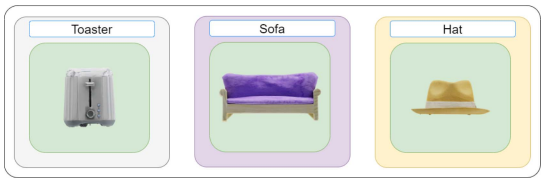


Fig. 9. Front-view used in experimental fixed object setting.

The study comprised four object interaction tasks: three *fixed objects* (hat, sofa, toaster; see Fig. 9) and one *free-choice object* selected by the participant. The fixed-object phase ensured comparability across users, while the free-choice task allowed open-ended exploration, testing Imagin3D's adaptability to diverse inputs. Only a frontal image was used per object, paired with a fixed descriptive prompt to standardize the initial conditions.

For each object, participants followed the same workflow: (i) providing a textual description and selecting a single 2D frontal image, from which the system generated multi-view images. (ii) After the Multi-view Generation took place and the MVQA provided detailed information about the presence or absence of each requested detail per-view, the user reviewed them. (iii) When inconsistencies are detected, users entered the SGI stage and proceeded to multi-view inpaint the views (for fixed objects, we instruct users to produce similar and reproducible inpainting actions). (iv) Once satisfied, participants reconstructed the final 3D object and explored it using the 3D viewer. (v) After each generation, participants completed a questionnaire, following a pairwise comparison protocol, to assess satisfaction, fidelity, and coherence at every stage of the pipeline. It is worth noting that, while this experimental setting was adopted to evaluate all the components in Imagin3D with a sequential approach, the users exploited the multi-step nature of our system by performing multi-step refinements in all defined settings, showcasing the system's "back-and-forth" approach (see Appendix B-A for more details, available online). Upon completing all object interactions, participants move on to a post-questionnaire. This final survey gathers comparative feedback regarding usability, cognitive load, and adoption willingness with the Imagin3D tool. Each user spent an average of 49.69 ($\pm$ 66.79) seconds to perform the SGI process. This preliminary results indicates acknowledge the usability of the Imagin3D interface. These results were calculated by exploiting data obtained through our logging system (more details in Appendix C-D, available online).

C. Measurements

The measured variables consist of explicit scale metrics delineating an evaluation framework inspired by several related works [40], [41], [42]. We introduced a novel combination of questionnaire-based surveys to measure the perception of users towards the generated 2D and 3D content (all designed as a 5-point Likert Scale, more details in Appendix A, available online). These measures were applied after each experimental phase, considering a within-subject setting, as reported in Section V-B. We evaluated the system through four dedicated constructs and standardized usability scales. 2D Image Perception focused on user engagement and satisfaction with the generated views (2DI-X) [42], while MLLM-Assisted Evaluation assessed the utility of language model guidance in supporting 2D generation (LAE-X). Guided Inpainting measured the effectiveness of HITL inpainting in improving results (INPED-X), and 3D Object Perception evaluated engagement, fidelity, satisfaction, and coherence of the 3D models (3D-X). In addition, we adopted established usability scales (SUS, NASA-TLX, and TAM) to capture overall usability and acceptance [43] (complete details on each construct and item are reported in the Appendix A-A, available online). For SUS, we provide scores for positively and negatively worded items to offer a more diagnostic interpretation of usability across conditions, following previous works [20], [44]. We highlight that Imagin3D is not a traditional DL pipeline in which one component can be removed while still yielding an informative output. The system works through the pipelining of its modules. Removing any one of these modules breaks the workflow. For this reason, the constructs placed in our experimental setting acts as a user-centered ablation study: we compared performance and perceptions when participants interacted with or without one of these modules, exploiting the construct described. Comparing 2DI and INPED constructs (Table II) isolates the contribution of interactive inpainting. The LAE construct directly assesses the value of MVQA guidance. Along with the final 3D perception scales, these evaluations allowed us to quantify the marginal contribution of each module.

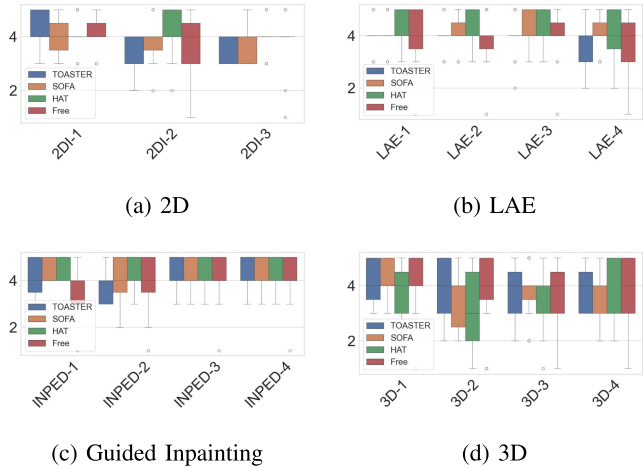
TABLE II
PRE/POST GENERATIVE INPAINTING VIEW SCORES COMPARISON (MEAN ($\pm$ STD)) ASSISTED BY THE MLLM

Viewpoint	Phase	HAT	SOFA	TOASTER	Free
Front	PRE	3.733 (0.854)	3.733 (0.998)	3.867 (1.024)	**4.2 (0.653)**
	POST	**4.6 (0.49)**	**4.267 (0.68)**	**3.933 (0.772)**	4.2 (1.046)
Front-Right	PRE	3.733 (0.854)	3.8 (0.909)	3.6 (0.879)	3.733 (1.062)
	POST	**4.533 (0.499)**	**4.067 (0.854)**	**4.2 (0.653)**	**4.067 (0.998)**
Right	PRE	3.333 (0.699)	3.467 (0.806)	3.333 (0.596)	3.533 (1.147)
	POST	**4.267 (0.772)**	**4.0 (1.033)**	**4.2 (0.653)**	**3.8 (1.108)**
Back	PRE	3.867 (0.718)	3.333 (0.699)	3.267 (1.236)	3.2 (1.108)
	POST	**4.6 (0.49)**	**3.4 (0.8)**	**3.4 (0.952)**	**3.467 (1.147)**
Left	PRE	3.467 (0.718)	3.467 (0.718)	**3.867 (0.618)**	3.533 (1.147)
	POST	**4.267 (0.772)**	**3.867 (1.024)**	3.8 (0.653)	**4.0 (1.155)**
Front-Left	PRE	3.867 (0.806)	3.667 (0.943)	3.533 (0.884)	3.733 (1.062)
	POST	**4.533 (0.499)**	**3.933 (0.998)**	**3.733 (0.772)**	**3.933 (0.998)**

Bold values indicate the higher score between the PRE and POST inpainting phases for each viewpoint-object combination.

D. Results

In this Section, we will describe the statistical framework that we followed and the obtained results. To analyze the obtained results, we performed descriptive statistics and data visualization. Specifically, we employed box plots on the main constructs grouped by each of the four considered objects on 2DI, LAE, INPED, 3D, and Usability constructs. Then, we performed a statistical analysis to analyze meaningful patterns in our observed data. First, we computed Cronbach's alpha to evaluate internal consistency and reliability. All constructs achieved a Cronbach's



(a) 2D (b) LAE

(c) Guided Inpainting (d) 3D

Fig. 10. Boxplots for the considered question items.

alpha above the commonly accepted threshold of $\alpha \geq 0.7$, indicating good reliability and supporting their inclusion in the analysis [45]. Then, we evaluated meaningful differences in all the examined constructs (i.e., 2DI, LAE, INPED, 3DP) through statistical tests. We first compared the scores obtained for all the constructs, comparing the different selected objects (fixed vs free ones). This analysis aims to assess whether some objects or modalities yield better results than others. We then compare the scores given by our users, comparing the generated views and the inpainting ones, to evaluate whether the inpainting views yield more attached and user-requested images. To ensure the validity of these comparisons, we performed the Wilcoxon signed-rank test, which is a non-parametric test designed for within-subject comparisons [46] (all variables were non-normal). This test determines whether a significant difference exists between two conditions. If a significant difference is detected ($p \leq 0.05$), only the direction of the effect (greater or lower) is reported. Conversely, if no significant difference is observed, the test statistic will not be included. Lastly, findings from the **User Attitudes and Usability** study are presented, incorporating both implicit and explicit measures. All the statistical analyses were performed using Python (v3.8) with the packages Matplotlib (v3.8.4), Pandas (v2.2.1), Pingouin (v0.5.5), Seaborn (v0.13.2), Scipy (v1.12.0), and Numpy (v1.26.4).

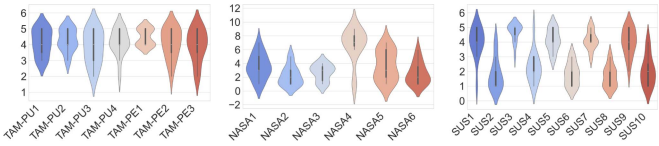
1) Statistical Analysis of Constructs: The boxplots presented in Fig. 10 regard the LAE, 2D, 3D, and INPED score distributions across all the experimental conditions (i.e., TOASTER, SOFA, HAT, Free). The median scores remain relatively high, clustering around values between 3 and 4, indicating consistent performance across different experimental conditions.

Regarding **2D constructs**, a significant difference was found only for item 2D-2 ("The generated views reflect my initial idea and request of object"), where HAT outperformed TOASTER ($p = 0.0258$, $stat = 9.0$) and marginally outperformed Free ($p = 0.0999$, $stat = 12.0$). No other 2D constructs showed statistical differences, suggesting that engagement and perceived robustness remain consistent across modalities, including the unconstrained Free condition. For **LAE constructs**, results highlight varied perceptions of MLLM assistance. While LAE-1 ("The MLLM guidance helped me identify missing objects in 2D content") showed no significant differences, high average scores across all conditions indicate consistent perceived usefulness. In LAE-2 ("The furnished scores provided by the LLM were useful in improving my 2D content"), HAT was significantly more beneficial than Free ($p = 0.0196$, $stat = 0.0$) and marginally better than TOASTER ($p = 0.0588$, $stat = 4.0$). For LAE-3 ("I felt more confident in completing the 2D generation with MLLM assistance"), structured conditions HAT and SOFA both outperformed Free ($p = 0.0339$, $stat = 0.0$), and SOFA outperformed TOASTER ($p = 0.0348$, $stat = 4.0$). Finally, LAE-4 ("The MLLM guidance allowed me to focus more on creativity rather than technical corrections") showed no significant differences; high overall scores suggest the MLLM successfully facilitates creativity regardless of condition. Concerning **3D constructs**, 3D-1 ("I was engaged with the 3D object at hand") and 3D-4 ("My modifications on generated views reflect my initial idea and request of object") yielded no significant differences. However, for 3D-2 ("I was satisfied with this type of generated 3D model"), HAT was rated significantly lower than TOASTER ($p = 0.0458$, $stat = 14.0$) and marginally lower than Free ($p = 0.0561$, $stat = 12.0$), while TOASTER outperformed SOFA ($p = 0.0751$, $stat = 8.0$). Regarding 3D-3 ("The realism of the 3D models helps to enhance my creativity and skills"), SOFA was rated significantly higher than HAT ($p = 0.0348$, $stat = 4.0$). Overall, while specific satisfaction and realism ratings varied, user engagement remained consistently high across all 3D conditions.

2) Statistical Analysis Pre and Post Inpainting: To evaluate whether the inpainting views yielded images more aligned to user requests, we directly compared the scores given by our users before (PRE) and afterwards (POST) the inpainting phase. Table II reports this comparison.

The scores for inpainting processes during *POST* increased in general relative to *PRE*, pointing to the fact that inpainting aided the process to enhance view quality. There are further variations of effectiveness among views in terms of inpainting. The top two *POST* scores are for the Front and *Back* for the *HAT* object (4.6 (0.49) in both cases). On the contrary, the side views (*Right, Left, Front-Right, Front-Left*) exhibit a more moderate ability to improve inpainting. The *Free* view, an unconstrained view, gets a good share of the highest *PRE* results, although it lacks marked improvements in *POST*.

To check whether these observations were statistically significant, we performed the Wilcoxon tests again on those variables. For HAT, significant differences were found between the 2DI and INPED conditions in the following: Front ($p = 0.0035$), Front-Right ($p = 0.0057$), Right ($p = 0.0062$), Back ($p = 0.0023$), Left ($p = 0.0057$), and Front-Left ($p = 0.0039$). Again, INPED received generally higher values. Across all views, there were no significant differences in the perception of quality between conditions for SOFA. In TOASTER, there were significant differences for Front-Right ($p = 0.0304$) and Right ($p = 0.0030$), suggesting that INPED was rated higher than 2DI. No significant differences were noted for Free across all views. These findings suggest that structured guidance under INPED conditions led to higher perceived quality for multiple



(a) TAM PU & PE (b) NASA (c) SUS

Fig. 11. Violin plots for questionnaire responses.

views, particularly for HAT and TOASTER, whereas SOFA and Free conditions showed no significant improvement. Future research should explore the specific elements within INPED that contribute to increased perceived quality and whether this trend extends to different object categories and task settings.

3) Usability Analysis: Regarding the Technology Acceptance and Usability, we here report the scores obtained from all the users in our post-questionnaires, where we furnished the SUS, NASA, and TAM scales, as mentioned in Section V. The results are visually depicted in Fig. 11.

These plots show consistently high TAM scores, reflecting positive perceptions of usefulness and ease of use, while NASA responses were more varied, particularly for workload measures, but overall indicated a low task load. Finallyh, SUS results revealed low scores for negative items and high scores for positive ones, confirming good usability. However, to uncover which factors may have lowered the usability of our system, we collected and conceptually analyzed the answers obtained from our qualitative study. Users appreciated how the system could produce 3D items quickly and found the integration of MLLM's key to assist users in object modifications. However, some users outlined how the modification of objects, in particular, the mask generation process, could be slightly tedious (a specific user study employing our log data is discussed in the Appendix C-D, available online). Finally, users suggest incorporating pop-up guidance and a history feature to improve user control over 3D model visualization.

VI. Discussions and Conclusions

In this work, we introduced Imagin3D, a novel system that leverages GenAI paradigms within an innovative HITL framework to enhance human-AI collaboration for 3D generation from a single image. In the system, we created the novel MVQA module, powered by MLLMs, and a consistency-aware inpainting mechanism to support users into a more accurate and context-aware 3D object generation. Furthermore, we designed and developed an intuitive UI that streamlines the entire process, allowing end users to interact with the system seamlessly. To evaluate our contributions, we conducted a quantitative analysis of the MVQA module on the GSO dataset, demonstrating its effectiveness in improving multi-view prompt adherence for user generation guidance. Additionally, we performed a user study to assess Imagin3D's ability to assist users in the 3D generation process. Our findings suggest that Imagin3D can effectively bridge the gap between AI-generated outputs and user intent, mitigating common GenAI controllability limitations in 3D Generation.

Nevertheless, different limitations have emerged. First, Imagin3D is currently designed and optimized for object-level generation, where canonical viewpoints provide adequate coverage. While this suffices for single objects or moderately complex multi-part structures, full scene-level generation remains challenging. Moreover, the MVQA component relies on prompt-conditioned question generation rather than image-conditioned ones. This ensures independence between generation and evaluation, while exposing the system to natural variability in user inputs. Nevertheless, hybrid approaches that combine textual and image-conditioned evaluation may improve system adaptability in some specific cases, such as novice users. Such extensions will also include features from recent advancements on multi-view captioning, adaptive view selection, and 3D VQA [47], [48], to improve viewpoint relevance and enable reasoning over generated content. Moreover, adopting classic text-conditioned inpainting diffusion models may not work well when the newly addedportion involves complex meshes. In this context, the guidance provided by image-conditioned Diffusion Models like ControlNet could improve structural fidelity, when paired with Conditions like Canny Edge Maps [49]. In our case, we chose not to adopt ControlNet-like models due to the fact that the requirements for additional visual conditions may limit the flexibility of the tool (in particular for areas within the mask itself), and so we opted for a trade-off between structural quality generation and diversity by adopting a simple Mask-Conditioned Diffusion Model. Future iterations of Imagin3D will see the integration of these techniques in order to explore their impact on the final generation. We also plan on expanding the visualization layer for the generated 3D model, in order to extend interactions, by integrating well-established libraries like Blender Python and PyMeshLab. Moreover, a natural extension of our framework involves full scene-level generation and editing, leveraging intelligent region selection through 3D segmentation, with multi-view and 3D inpainting for visual refinement. Finally, we will conduct broader user studies across diverse application domains to refine Imagin3D and assess its impact in different scenarios.

References

[1] K. Zhao and A. Larsen, "Challenges and opportunities in 3D content generation," 2024, *arXiv:2405.15335*.
[2] S. Dong, L. Ding, Z. Huang, Z. Wang, T. Xue, and D. Xu, "Interactive3D: Create what you want by interactive 3D generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 4999–5008.
[3] P. Li et al., "ERA3D: High-resolution multiview diffusion using efficient row-wise attention," 2024, *arXiv:2405.11616*.
[4] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3D using 2D diffusion," 2022, *arXiv:2209.14988*.
[5] L. Stacchio, E. Balloni, L. Gorgoglione, M. Paolanti, E. Frontoni, and R. Pierdicca, "X-NR: Towards an extended reality-driven human evaluation framework for neural-rendering," in *Proc. Extended Reality: Int. Conf., XR Salento 2024*, Lecce, Italy, 2024, pp. 305–324.
[6] F. Faruqi et al., "Style2Fab: Functionality-aware segmentation for fabricating personalized 3D models with generative AI," in *Proc. 36th Annu. ACM Symp. User Interface Softw. Technol.*, 2023, pp. 1–13.
[7] N. Pangakis and S. Wolken, "Keeping humans in the loop: Human-centered automated annotation with generative AI," 2024, *arXiv:2409.09467*.
[8] Y. He et al., "LLMs meet multimodal generation and editing: A survey," 2024, *arXiv:2405.19334*.
[9] I. Carnat, G. Comandé, D. Licari, and C. De Nigris, "From human-in-the-loop to LLM-in-the-loop for high quality legal dataset," *i-lex*, vol. 17, no. 1, pp. 27–40, 2024.

[10] J. Cho et al., "Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-image generation," 2023, *arXiv:2310.18235*.

[11] C. Li et al., "Generative AI meets 3D: A survey on text-to-3D in AIGC era," 2023, *arXiv:2305.06131*.

[12] H. Gani, S. F. Bhat, M. Naseer, S. Khan, and P. Wonka, "LLM blueprint: Enabling text-to-image generation with complex and detailed prompts," 2023, *arXiv:2310.10640*.

[13] Y. Hong, C. Lin, Y. Du, Z. Chen, J. B. Tenenbaum, and C. Gan, "3D concept learning and reasoning from multi-view images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 9202–9212.

[14] X. Pan, A. Tewari, T. Leimkühler, L. Liu, A. Meka, and C. Theobalt, "Drag your gan: Interactive point-based manipulation on the generative image manifold," in *Proc. ACM SIGGRAPH 2023 Conf. Proc.*, 2023, pp. 1–11.

[15] A. Haque, M. Tancik, A. A. Efros, A. Holynski, and A. Kanazawa, "Instruct-NeRF2NeRF: Editing 3D scenes with instructions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 19740–19750.

[16] E. Weber et al., "NeRFiller: Completing scenes via generative 3D inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 20731–20741.

[17] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang, "NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction," 2021, *arXiv:2106.10689*.

[18] F. Lin, A. Z. Wang, M. D. Rahman, D. A. Szafir, and G. J. Quadri, "Striking the right balance: Systematic assessment of evaluation method distribution across contribution types," in *Proc. 2024 IEEE Eval. Beyond-Methodological Approaches Visual.*, 2024, pp. 129–135.

[19] Y. Xiu, J. Chilukuri, S. Sen, and M. Gorlatova, "Say it, see it: A systematic evaluation on speech-based 3D content generation methods in augmented reality," 2025, *arXiv:2508.12498*.

[20] E. Santarnecchi et al., "MineVRA: Exploring the role of generative AI-driven content development in XR environments through a context-aware approach," *IEEE Trans. Vis. Comput. Graph.*, vol. 31, no. 5, pp. 3602–3612, May 2025.

[21] M. Behravan, M. Haghani, and D. Gračanin, "Transcending dimensions using generative AI: Real-time 3D model generation in augmented reality," in *Proc. Int. Conf. Hum.- Comput. Interact.*, 2025, pp. 13–32.

[22] Y. Hu et al., "TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 20406–20417.

[23] N. Ravi et al., "SAM 2: Segment anything in images and videos," 2024, *arXiv:2408.00714*.

[24] Y. Hao, Z. Chi, L. Dong, and F. Wei, "Optimizing prompts for text-to-image generation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 66923–66939.

[25] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10684–10695.

[26] L. Downs et al., "Google scanned objects: A high-quality dataset of 3D scanned household items," in *Proc. 2022 Int. Conf. Robot. Automat.*, 2022, pp. 2553–2560.

[27] R. Liu, R. Wu, B. Van Hoorick, P. Tokmakov, S. Zakharov, and C. Vondrick, "Zero-1-to-3: Zero-shot one image to 3D object," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 9264–9275.

[28] M. Liu et al., "One-2-3-45: Any single image to 3D mesh in 45 seconds without per-shape optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 22226–22246.

[29] Y. Liu et al., "SyncDreamer: Generating multiview-consistent images from a single-view image," 2023, *arXiv:2309.03453*.

[30] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.

[31] S. Banerjee and A. Lavie, "Meteor: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summarization*, 2005, pp. 65–72.

[32] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proc. Text Summarization Branches Out*, 2004, pp. 74–81.

[33] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 382–398.

[34] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "CLIPScore: A reference-free evaluation metric for image captioning," in *Proc. 2021 Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 7514–7528.

[35] X. Wang, D. Song, S. Chen, C. Zhang, and B. Wang, "LongLLaVA: Scaling multi-modal LLMs to 1000 images efficiently via hybrid architecture," 2024, *arXiv:2409.02889*.

[36] H. Touvron et al., "Llama: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.

[37] X. Long et al., "Wonder3D: Single image to 3D using cross-domain diffusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 9970–9980.

[38] V. Chheang et al., "Towards anatomy education with generative ai-based virtual assistants in immersive virtual reality environments," in *Proc. IEEE Int. Conf. Artif. Intell. Extended Virtual Reality*, 2024, pp. 21–30.

[39] P. Salomoni, C. Prandi, M. Roccetti, L. Casanova, L. Marchetti, and G. Marfia, "Diegetic user interfaces for virtual environments with HMDs: A user experience study with oculus rift," *J. Multimodal User Interfaces*, vol. 11, pp. 173–184, 2017.

[40] Y. Jin, M. Chen, T. Goodall, A. Patney, and A. C. Bovik, "Subjective and objective quality assessment of 2D and 3D foveated video compression in virtual reality," *IEEE Trans. Image Process.*, vol. 30, pp. 5905–5919, 2021.

[41] Y. Nehmé, F. Dupont, J.-P. Farrugia, P. Le Callet, and G. Lavoué, "Visual quality of 3D meshes with diffuse colors in virtual reality: Subjective and objective evaluation," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 3, pp. 2202–2219, Mar. 2021.

[42] M. C. Fink, D. Sosa, V. Eisenlauer, and B. Ertl, "Authenticity and interest in virtual reality: Findings from an experiment including educational virtual environments created with 3D modeling and photogrammetry," in *Proc. Front. Educ.*, 2023, Art. no. 969966.

[43] L. Longo, "Subjective usability, mental workload assessments and their impact on objective human performance," in *Proc. IFIP Conf. Hum.-Comput. Interact.*, 2017, pp. 202–223.

[44] S. Hajahmadi, L. Stacchio, A. Giacché, P. Cascarano, and G. Marfia, "Investigating extended reality-powered digital twins for sequential instruction learning: The case of the Rubik's cube," in *Proc. 2024 IEEE Int. Symp. Mixed Augmented Reality*, 2024, pp. 259–268.

[45] M. Tavakol and R. Dennick, "Making sense of Cronbach's alpha," *Int. J. Med. Educ.*, vol. 2, 2011, Art. no. 53.

[46] B. Rosner, R. J. Glynn, and M.-L. T. Lee, "The Wilcoxon signed rank test for paired comparisons of clustered data," *Biometrics*, vol. 62, no. 1, pp. 185–192, 2006.

[47] T. Luo, C. Rockwell, H. Lee, and J. Johnson, "Scalable 3D captioning with pretrained models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 75307–75337.

[48] T. Luo, J. Johnson, and H. Lee, "View selection for 3D captioning via diffusion ranking," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 180–197.

[49] C. Chen, C. Nguyen, T. Groueix, V. G. Kim, and N. Weibel, "Memovis: A genai-powered tool for creating companion reference images for 3D design feedback," *ACM Trans. Comput.- Hum. Interact.*, vol. 31, no. 5, pp. 1–41, 2024.

[50] P. Esser et al., "Scaling rectified flow transformers for high-resolution image synthesis," in *Proc. 41st Int. Conf. Mach. Learn.*, 2024, pp. 12606–12633.

[51] D. Giunchi, N. Numan, E. Gatti, and A. Steed, "DreamCodeVR: Towards democratizing behavior design in virtual reality with speech-driven programming," in *Proc. 2024 IEEE Conf. Virtual Reality 3D User Interfaces*, 2024, pp. 579–589.

[52] C.-H. Yeh et al., "Seeing from another perspective: Evaluating multi-view understanding in MLLMs," 2025, *arXiv:2504.15280*.

Open Access funding provided by 'Università Politecnica delle Marche' within the CRUI CARE Agreement