Inceptor: Automated Generation of Social Scenarios in Virtual Reality

Dan Pollak
Reichman University
Herzliya, Israel
dandan888@gmail.com

Ranit Maimon
Reichman University
Herzliya, Israel
ranitnuna@gmail.com

Maya Shekel
Reichman University
Herzliya, Israel
shekelmaya@gmail.com

Jonathan Giron
Reichman University
Herzliya, Israel
giron.jonathan@gmail.com

Friedman Doron*
Reichman University
Herzliya, Israel
doronf@runi.ac.il

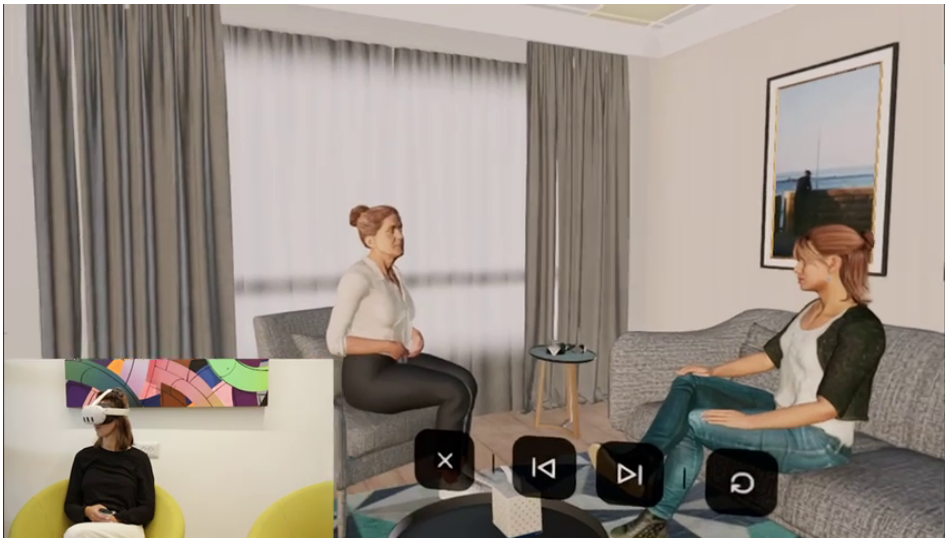

Figure 1: A snapshot from a scene automatically generated by the Inceptor, experienced in VR.

Abstract

Inceptor is a novel software tool designed to expedite the development of social virtual reality (VR) scenarios. It enables users, particularly non-experts, to effortlessly transform text scripts into complex 3D/VR environments featuring multiple virtual humans, all within Unity. The tool operates across three levels of abstraction—the text script, the scenario data structure, and the Unity scenes—providing users with unprecedented flexibility to modify scenarios at any developmental stage. Unlike end-to-end generative models, *Inceptor* incorporates a human-in-the-loop approach, ensuring enhanced control and customization of outputs.

The application of *Inceptor* spans from basic research to practical applications within the psychological domain. This paper presents an evaluation of *Inceptor* in a project focused on resilience training in mental health contexts. Using the automated pipeline allowed us to generate 32 VR episodes in a very short time, spanning more than two hours of carefully designed content. Additionally, we assessed the feasibility of using large language models (LLMs) to autonomously select appropriate animations. The content is now being used for training diverse professional groups in mental health. Encouraged by positive feedback and early evidence of its utility, *Inceptor* will be released as an open-source project, poised for widespread adoption and community-driven enhancements.

*Corresponding author.

IVA '25, Berlin, Germany

ACM ISBN 979-8-4007-1508-2/25/09
https://doi.org/10.1145/3717511.3747061

CCS Concepts

• **Computing methodologies** → **Virtual reality**; *Natural language processing*; • **Human-centered computing** → **Interaction design**; *User interface toolkits*; *Mixed / augmented reality*; • **Applied computing** → *Computer-assisted instruction.*

Keywords

Virtual Reality, Social VR, Scenario Generation, Large Language Models, Virtual Humans, Resilience Training, Interactive Storytelling

ACM Reference Format:

Dan Pollak, Ranit Maimon, Maya Shekel, Jonathan Giron, and Friedman Doron. 2025. Inceptor: Automated Generation of Social Scenarios in Virtual Reality. In *ACM International Conference on Intelligent Virtual Agents (IVA '25), September 16–19, 2025, Berlin, Germany.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3717511.3747061

1 Introduction

Authoring believable social interactions in virtual reality (VR) traditionally requires weeks of custom animation, dialogue scripting, and engine integration. *Inceptor* [22] streamlines this process: given a text script, it automatically generates an interactive Unity scene featuring multiple virtual humans with voiced dialogue and non-verbal behavior. While end-to-end text-to-3D generation systems are emerging [14, 23], they still lack the fidelity needed for generating complete simulations. Thus, *Inceptor* adopts a *human-in-the-loop* approach, automating routine tasks while letting designers fine-tune staging and timing within the familiar Unity editor—combining generative speed with creative control.

Unlike the common *agent-first* paradigm, where autonomous characters drive emergent narratives, *Inceptor* follows a *story-first* approach: authors start with a cohesive, pre-planned screenplay, and the engine ensures all character actions serve that narrative. This guarantees that training or educational goals are met without characters deviating from the intended storyline.

Our contributions are threefold. First, we present *Inceptor*, an open-source pipeline [1] that converts linear and branching dialogue scripts into complete social-VR scenes with embodied virtual humans. Second, we evaluate the use of large language models to recommend body-language animations, showing that automated suggestions can significantly reduce authoring effort. Finally, we detail a real deployment: a 32-scene resilience training program (15 skills, about 2 hours of content) produced in roughly one week. For a video see [2].

2 Background

Recent advances in deep neural networks (DNNs) have expanded the automation of 3D model creation, game logic, and procedural content—often driven by textual descriptions. Generative models such as Generative Adversarial Networks (GANs) [21, 33], Variational Autoencoders (VAEs) [20, 36], and diffusion models [32] have been central to this progress. These methods have been applied across domains—from generating 3D human figures [15] and animated landscapes [2] to building complex objects from simple text [5]. Differentiable rendering combined with explicit, implicit, and hybrid scene representations enables detailed 3D content consistent with textual inputs [19, 35]. Complementary advances in high-fidelity facial animation [4], text-to-gesture mapping [3, 10], and text-to-3D texture generation [34] further demonstrate the transformative impact of generative models on content workflows.

[1] https://github.com/Advanced-Reality-Lab/inceptor
[2] https://tinyurl.com/2mhydfm9

Beyond the generation of specific assets, several studies addressed the challenge of creating a complete interactive 3D/VR environment. Riedl and Young [25] discuss tensions between narrative structure and character intentionality in automated story generation. While their focus is on algorithmically producing coherent narratives, we make a similar distinction but focus on automated 3D/VR scene generation. We distinguish between two paradigms in interactive systems: *story-first* and *agent-first*. The former, exemplified by *Inceptor*, begins with a fixed, pre-authored script emphasizing authorial control over plot, actions, and dialogue. Such narrative-centric design suits training and entertainment contexts, where coherent, predetermined storylines guide interactions and ensure controlled, meaningful experiences [27]. The latter centers on autonomous agents whose behaviors generate *emergent* narratives without explicit scripts, arising through simulation.

Several studies have facilitated interactive game narratives, including Expressionist [26], StoryAssembler [9], IFDBMaker [31], StoryPlaces [13], Ensemble Engine [28], Comme il Faut [18], and Villanelle [17]. Li et al. [16] proposed converting free text into structured JSON for scene creation in Autodesk Maya, covering skyboxes, 3D models, scales, positions, code, and animation cues.

More recently, some projects have leveraged progress in DNNs: Chen et al. [7] developed a VR storytelling tool mapping events to visuals and animations via action triggers. Beinema et al. [6] presented Agents United, an open platform for multi-agent conversational systems relying on emergent narratives, contrasting with our story-first method. Recent works integrate large language models for scene generation, including Guinchi et al.'s VR speech-to-code platform [11], text-to-avatar creation [12], and multi-agent animated content generation [29].

The *Inceptor* framework introduces a novel LLM-based pipeline for VR authoring, designed to integrate future deep neural network components (e.g., end-to-end voice and animation generation), allowing advances in generative modeling to complement rather than replace our narrative-driven approach.

3 Method

3.1 Design Principles and Goals

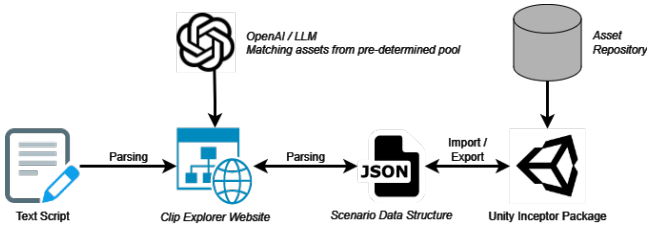

Figure 2: *Inceptor* schematic diagram.

Inceptor is guided by four design principles:

(1) **Intuitive for diverse authors:** Designed for users with little VR experience to quickly create rich social scenarios, while enabling experienced Unity developers to automate routine tasks and focus on creative production.

(2) **Support for iteration and modification:** Three abstraction layers maintain automatic consistency, allowing flexible edits without losing automation benefits.
(3) **Focus on social scenarios:** Prioritizes virtual humans with verbal and non-verbal communication, targeting large-scale or repeated structured episodes rather than unique one-off experiences.
(4) **Scalability and extensibility:** Built to accommodate future extensions and diverse applications.

Inceptor employs a three-layer approach to create immersive VR scenarios (Figures 2, 3):

(1) **Cinematic Script Layer:** A natural-language movie-script style layer enabling authors to draft scene narratives with dialogue and basic actions for any number of actors.
(2) **Scenario Data Structure:** Transforms narratives into a cinematic data structure optimized for multi-character social interactions, including spatial and temporal details. A web-based UI supports manual editing.
(3) **Unity Project:** Mirrors the scenario data with Unity-specific elements and is accessible for human editing via an *Inceptor* plugin.

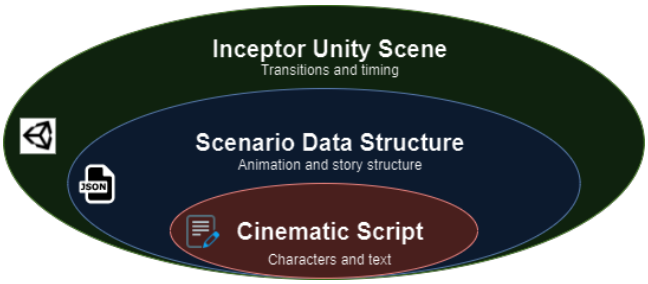

Figure 3: The conceptual structure of the *Inceptor* framework. Each layer has a higher level of abstraction and lower resolution, missing some spatial and temporal details.

Facial expressions, body movements, and eye gaze are added incrementally, either via generative AI (default GPT-4, see next section) or manual adjustment. This hybrid approach enhances flexibility and productivity.

The Unity scene reflects the scenario data but adds details like animation transitions, preset by default and customizable by experts for precise interaction and movement control. This ensures *Inceptor* supports both automation and high-quality, detailed productions.

Following a "story-first" approach, the cinematic script remains the central element, encoding predefined character information rather than focusing on autonomous agents.

3.2 Phase I: Parsing Free Text to a Data Structure

First, the user creates a text file resembling a movie script. We illustrate this with a simple example from the Resilience Hub project (Section 4.4), showing a dialogue between 'Rachel' and 'Jane', a patient and a psychologist (Fig. 4).

Inceptor first parses the script into a JSON data structure, breaking it into a sequence of "clips" — each representing a spoken line by a character. The clip also encodes non-verbal behaviors of non-speaking characters. For example, the line 'Rachel: "I've just had the most terrible week"' is mapped to Rachel as the speaker.

```
Rachel: "I've just had the most terrible week, I feel bad"
Jane: "Let's talk about this feeling and try to define it better"
Jane: (Leaning Backwards): "What did you mean by it was a terrible
week?"
Rachel: "I was so angry at my kids and my customers"
Rachel: "One of my customers was late and I just yelled at her"
```

Figure 4: An example of a text script, serving as input to *Inceptor*.

This parsing runs automatically and supports batch processing. Human editing is possible through an interactive web tool called the "Clip Explorer", enabling manual adjustments.

Clip Name	Text	Character	is Speaker?	Animation Type	Animation Choice
Clip 0	"I've just had the most terrible week, I feel so bad"	Noa	TRUE	Mood	Disturbed
				Reaction	Upset
		Nirit	FALSE	Mood	Listen Confident
				Reaction	Nod
Clip 1	"Let's Talk about this feeling and try to define it better"	Noa	FALSE	Mood	Annoyed
				Reaction	Upset
		Nirit	TRUE	Mood	Interested
				Reaction	Faint Smile
Clip 2	"What did you mean by it was a terrible week?"	Noa	FALSE	Mood	Disturbed
				Reaction	Impatient
		Nirit	TRUE	Mood	Listen Confident
				Reaction	Friendly
Clip 3	"I was so angry at the kids and my customers."	Noa	TRUE	Mood	Listen And Think
				Reaction	Irritated
		Nirit	FALSE	Mood	Comfortable
				Reaction	Faint Smile
Clip 4	"One of my customers were late and I just yelled at her"	Noa	TRUE	Mood	Annoyed
				Reaction	Impatient
		Nirit	FALSE	Mood	Listen Confident
				Reaction	Nod

Figure 5: Sample script enriched with animation clips selected by the LLM.

Next, animation selection is performed. Script writers may include annotations or stage directions (e.g., 'leaning backward'). A large language model (LLM) uses these cues and the script context to recommend suitable animation clips (Fig. 5). Because animation choice is subjective, users can override LLM suggestions via the Clip Explorer. The system supports multiple LLMs for flexibility.

For human models and animation, *Inceptor* relies on a library of assets including the open-source Rocketbox[3]. Our tool includes 50 of the 115 Rocketbox rigged avatars representing diverse races, genders, ages, and occupations. It also contains several dozen Rocketbox animation files, and the library can be easily extended with more rigs and animations.

We also introduce "animation buckets"—predefined subsets of animations per project. For example, a scenario with seated characters can restrict selections to a "sitting" bucket, improving animation relevance and speed, and allowing users to filter out unwanted clips.

Since no single animation is "correct", creators often want to override LLM choices. While experts can replace animations directly in Unity, non-experts face barriers. Thus, the Clip Explorer enables manual review and editing of chosen animations, lowering technical hurdles.

[3]https://github.com/microsoft/Microsoft-Rocketbox

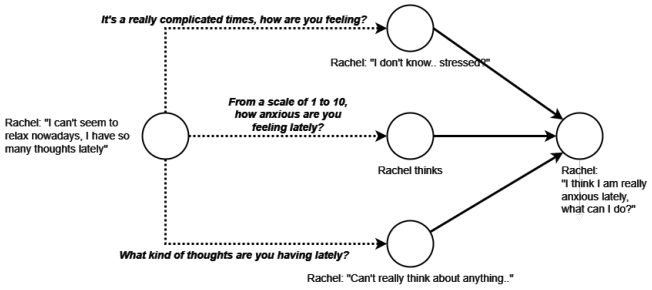

Figure 6: An interactive script example with a decision "fork", as displayed in the Clip Explorer.

3.2.1 Interactive Storytelling. In addition to converting linear scripts into social scenarios with virtual humans, our framework also supports non-linear, interactive scenarios. Our system focuses on social situations involving communication- both verbal and non-verbal. To support such interactive narratives, the Clip Explorer allows you to change the linear structure of a text script with "Fork" clips. A Fork clip allows the user to control the flow of the clips instead of having the system run them in a linear manner. The Inceptor design is such that each clip is self-contained. This allows defining arbitrary tree structures of clips using the Clip Explorer.

In Figure 4 we presented a simple linear fragment of a script. Next, we describe how the system supports interaction: in this modified example, we provided the participant with three options to choose from, and then the script moves to a specific clip accordingly in a question/response scheme (see fig. 6).

The example above was part of a training scenario, in which we deliberately limited the participant to one of three choices (see Section 4.4). Our approach is characterized as "story-first" so interaction is limited, but the mechanisms for interaction can be further extended and explored. As the clip explorer allows to define users to define a "next" clip for each clip, it allows for more complex story graph structures other than question/response format.

3.3 Phase II: Data Structure to Unity Scene

The next step involves converting the JSON structure representing the scene to Unity. We provide a Unity package that adds an Inceptor module into Unity. The module is split into two main parts: asset collection and scene creation.

Asset collection is a simple fetch protocol to import the essential files to the project, mainly the animation files, which can be imported from either a remote server or a local folder. It is also possible to add all assets into the project manually; we leave the asset provider as a different module as we expect each user to set their own asset repository.

The scene creation module receives an *Inceptor* JSON and builds the scene within Unity. In this phase, the system goes over all the selected assets, and prompts the user to replace assets manually or alert if they are missing. When completed, the system has created all required assets within the open scene in Unity. Since interactivity schemes can change among applications, we allow developers to provide and change "interaction schemes" into the wizard itself. Interaction schemes can be simple point-and-click user interfaces or more complex systems involving speech-to-text, LLMs, or game mechanics. As long as the interaction scheme provides to the *Inceptor* the decision in the fork, it is an independent module that can be changed and shared easily between different projects.

The *Inceptor* first serializes the JSON into a C-sharp object, and then converts it into components that connect with Unity‘s audio and animation components. To make the animations diverse and expressive, we use different animation layering for the body animations and facial animations that provide the designer with the ability to apply different animations on different body parts including the legs, torso, and face . Also, we introduce a distinction between ‘mood‘ and ‘reaction‘ animations – the first are relatively long clips of low frequency motion, and the latter are short term actions such as rolling your eyes in surprise, yawning, or similar. This allows us to play up to four animations at the same time, layering and customizing the animation for each author‘s specific need. We use Unity‘s state-machine Animator module in order to enable the user to customize transitions between each animation state.

Inceptor is equipped with capabilities for exporting and updating VR scenarios, ensuring a seamless workflow for users who iterate on their projects within Unity or utilize tools such as the Clip Explorer (2. This feature is particularly valuable when scenarios are dynamically edited or enhanced directly in the Unity environment. The system allows for updates and modifications to be re-imported into the original project without disrupting the custom settings that have been manually adjusted by users. These include the placement of characters within the scene, the timing of audio clips, and any specific delays that have been set. This ensures that enhancements or corrections to the scenario can be integrated smoothly, without overwriting critical custom parameters that have been meticulously set by the user. This functionality not only preserves the integrity of the original design but also enhances user efficiency by reducing the need to reapply customization after each update manually. By supporting such flexible update mechanisms, *Inceptor* fosters an iterative development process, encouraging continuous refinement and improvement of VR scenarios.

4 Results

4.1 Automated Production of the Resilience Hub

Resilience has emerged as a critical skill for maintaining well-being, especially in the face of global crises such as pandemics and conflicts [8]. Its relevance extends beyond extraordinary circumstances to everyday life, supporting individuals in coping with common stressors and setbacks [24]. Despite the recognized importance of resilience, widespread implementation remains challenging. Clinical practitioners, who are best equipped to impart resilience training, are in short supply. Consequently, one key strategy is to train a broader range of professionals —- including non-therapists —- to provide basic resilience skills. Yet practical strategies to develop these skills often stay confined to clinical settings, limiting access for the general population.

Resilience, as defined by the American Psychological Association, encompasses the capacity to adapt successfully in the face of adversity, trauma, or significant stress, relying on a blend of genetic, developmental, psycho-social, and psychological factors [1]. By reframing resilience as a "muscle" that can be strengthened, our VR-based training draws on principles from Cognitive-Behavioral Therapy (CBT) and Interpersonal Psychotherapy (IPT). The result is a more tangible, skill-based approach that can be accessible to a broader audience, potentially influencing mental health on a larger scale.

In response, we developed a novel resilience training paradigm that utilizes a VR platform to teach and practice resilience as a set of discrete, trainable skills. A clinical psychologist in training collaborated with an experienced Prof. of Clinical Psychology to transform theoretical principles into implementable scripts, drawing on existing tutorial videos and lectures. Together with the VR development team, they developed a fictional persona for the virtual patient in consultation with two mental health institutions via a user-centered design process. The virtual patient is portrayed as a mother of two, possibly divorced, working in the nail-care industry. Since undergoing a tragic event, she has faced multiple functional difficulties.

Based on resilience theory and practice, the psychologists identified 15 relevant skills in five categories: emotional, interpersonal, bodily, cognitive, and behavioral. Each skill was represented in two distinct scripts: an *instructional* sequence and a subsequent *interactive* sequence. In the instruction phase, participants watch a counselor discuss the virtual patient's challenges, placing the focal resilience skill in context. In the interactive phase, participants must actively apply the same skill by engaging with the virtual patient in conversation.

Inceptor was used to build a comprehensive program designed to enhance these 15 resilience skills. This program—comprising 32 scenarios (an introduction plus 15 pairs of scenes) —- relies on the same virtual patient and therapist throughout. Each "guidance" scene is a linear, non-interactive portrayal of the therapist demonstrating the skill with Rachel (Fig. 1, top), while each "practice" scene allows participants to interact with the virtual patient directly (Fig. 7).

In the interactive scripts, the virtual patient delivers a line of dialogue, after which participants must choose one of three potential responses. While all responses are generally acceptable, one consistently aligns best with the targeted resilience skill.

All scripts underwent extensive user testing, incorporating feedback from the partnering mental health facilities to refine content, improve engagement, and ensure clinical accuracy. For each script this consisted of a series of low-fidelity user testing, comprised of role playing with a human patient reading from a script.

Using *Inceptor* we produced an entire resilience–training curriculum that comprises **32 VR scenes**. For each of the **15 resilience skills** the designers wrote (i) a *linear guidance* scene in which a counsellor demonstrates the skill, and (ii) an *interactive practice* scene in which the trainee must apply that skill while conversing with a virtual patient. Two short introductory scenes—one that familiarises trainees with the VR interface and another that introduces the patient's back-story—bring the total to 32. On paper the curriculum spans approximately **100 pages of screenplay** and yields about **150 minutes** of spoken dialogue and non-verbal behaviour when rendered in VR.

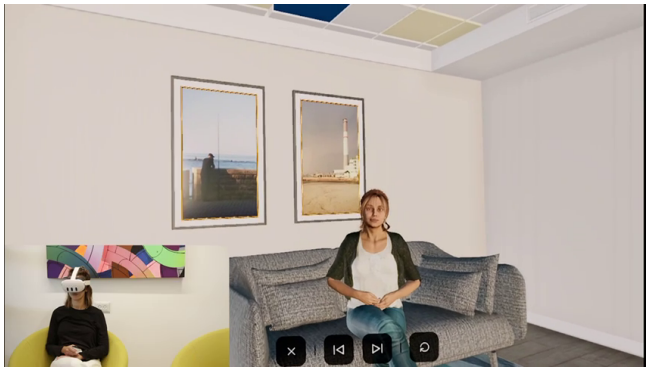

Figure 7: A snapshots from the resilience project, from the interactive scenario, in which the participant is facing the virtual patient.

The time-consuming part of the project lay *up-stream*: script writing, expert consultation, and several rounds of low-fidelity user testing with clinicians and field practitioners. However, once the scripts were determined, the *Inceptor* pipeline converted them into fully playable Unity projects in roughly one week, by a single Unity expert.

The only task that still required manual attention was the assignment of body-language animations—dozens of candidate clips for each utterance across 32 scenes composed of multiple clips. The lead animator initially preferred to hand-pick gestures and facial cues, skeptical of automatic choices. Yet, as we show in Section 4.2, even a simple LLM prompt produced animation selections that coincided with the expert's picks surprisingly often, indicating that future iterations of *Inceptor* could delegate most of this step to AI while still leaving experts in control of the final polish.

4.2 Automated Animation Retrieval

A key question during development was whether the LLM could autonomously handle the task of animation selection or if it would better function as a co-creative tool alongside the user. To evaluate this, we tested the LLM's ability to choose animations independently. Specifically, we utilized GPT-4o[4] to select animations for character behaviors.

To compare the LLM choices with human judgment, we used 4 cinematic scripts from the resilience-hub mental health project. The original animation selections by the screenwriter were treated as "Expert" choices, and we also asked a "Novice" user, with no prior animation experience, to independently select animations they deemed appropriate based on the script. As there is no "correct" solution, the goal was to evaluate the level of agreement among the three decisions makers ("AI" model, expert, and novice).

In total, 228 distinct animation choices were evaluated, with half representing "mood" animations and the other half "reaction" animations. The "mood" animations consisted of 15 different options, while the "reaction" animations included 28 options, including the possibility of "no reaction".

[4]https://openai.com/gpt-4o-contributions/

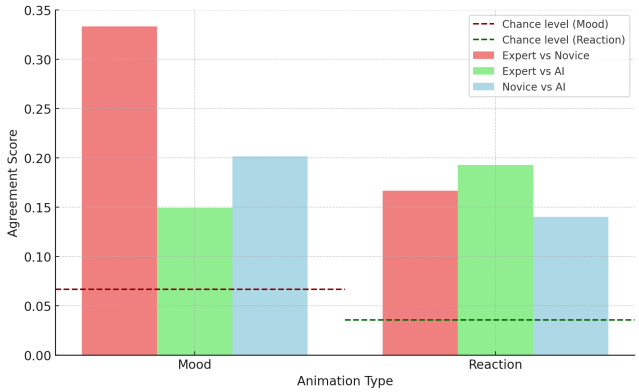

Figure 8: The similarity rate between all taggers, shown separately for mood and reaction. Chance level agreement is 0.067 for "mood" and 0.036 for "reactions".

Figure 8 shows the results, indicating above chance level but low agreement overall. In mood, expert and novice seemed to agree, more than both with the AI. However, the expert used most options (26), the AI next (28) and the novice the least (13). So the AI was more diverse than the novice, and almost as the expert.

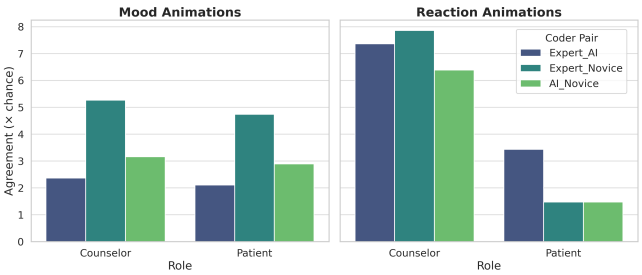

Figure 9: This figure presents the normalized agreement rates between three pairs of coders (Expert vs AI, Expert vs Novice, and AI vs Novice) for two types of animations—Mood and Reaction—across two characters, Rachel and Jane. The y-axis represents the normalized agreement rate, calculated by dividing the actual agreement rate by the chance level of selecting an animation (1/15 for Mood animations and 1/28 for Reaction animations). Each pair of bars represents a different character, with different colors indicating different pairs of coder comparisons. The normalization allows for a direct comparison of agreement above what would be expected by chance, highlighting the effectiveness of the AI and human coders in consistently selecting relevant animations for both mood and reaction contexts.

Figure 9 data delves deeper, revealing varying levels of agreement between coders for both mood and reaction animations. In the case of mood animations, the normalized agreement rates between the expert and AI are relatively low (around 2.1 to 2.4), but comparable to the agreement rates between AI and the novice coder (2.9). This suggests that the AI's performance in selecting appropriate mood animations is not significantly behind that of the novice coder and might indicate that the AI can replicate a novice-level understanding of appropriate animations for the given scenarios.

For reaction animations, the AI's performance relative to the expert varies greatly between the characters (3.44 for the patient and 7.37 for the counselor), suggesting that the AI's effectiveness might depend on the specific character context or the complexity of the reaction required. However, the AI consistently matches the novice's selections for the patient (1.47) and performs comparably for the counselor, which supports the notion that the AI is at least as good as a novice in selecting appropriate animations for reaction scenarios.

This evidence could be interpreted as supporting the inclusion of AI as a viable assistant in scenarios where novice-level performance is acceptable. The AI's consistent performance across characters and types of animations suggests that with further training and refinement, it could potentially reach or exceed the proficiency of more experienced coders. However, the higher agreement rates between the expert and the novice for the counselor in reaction animations (7.86) compared to those involving AI (1.47) indicate that there are still areas where human coders outperform AI, likely in more nuanced or complex interaction scenarios where human judgment is critical.

When also taking into account speaking vs listening behaviors we have very sparse data, but notably, Expert vs. Novice agreement rates are higher for mood animations, especially for the patient when talking. AI's performance aligns closely with that of the Novice in most scenarios, and in some cases (such as mood animations for the therapist), the AI and Novice show significant agreement.

Finally, average agreement rates across all conditions without any splits are as follows: Expert vs. AI: 17.11%, Expert vs. Novice: 25.00% and AI vs. Novice: 17.11%. These values confirm the previous observation that the agreement between the Expert and the Novice is higher than the agreements involving the AI.

4.3 Participant Feedback

The resilience training content is now being successfully deployed. A comprehensive evaluation study is beyond the scope of this paper, but we provide some anecdotes related to the design and implementation of the *Inceptor*.

4.4 Evaluation Study: Method

Participants: The study included a total of 28 participants aged 25-60 (21 females: 75%, M= 33.6, SD= 7.5), drawn from two primary groups: resilience coaches (n=17) and mental health professionals (n=11). The participants from the latter came from diverse sectors such as psychiatry, psychology, social work, and nursing.

Procedure: The evaluation was conducted in the target institutes, as part of mental health training sessions. During each pilot test, participants engaged with one of the VR resilience training modules, comprised of three chapters: i) Orientation, which provided an overview of the training structure and instructions on using the VR; ii) a skill-based training session focused on resilience techniques (Figure 1, top); and iii) the practice of the learned skill in a simulated interaction with the virtual agent "Rachel", a 35-year-old distressed mother who represented the patient (Figure 7). In

the interactive practice, the participants were instructed to respond to the virtual patient by reading out loud one of three options displayed on the screen, with real-time feedback provided based on their selected response.

The study was carried out during the development of the VR experience, and evaluated a preliminary version with a limited number of animation dataset to draw from. The study was conducted using Quest 2 and Quest 3 devices (Meta, inc).

The interviews were designed with open-ended questions focused on key areas, including the usability of the interface, realism of the virtual experience, the effectiveness of the simulation in practicing resilience skills and its contribution to participants' self-efficacy in applying those skills in real-world scenarios. For example, participants were asked:*"To what extent did the simulation feel like a real-world intervention? What factors contributed to or detracted from this?", "Do you feel the simulation could enhance your confidence in applying the resilience skills practiced and how?", and "How comfortable was your interaction with the system interface during the practice? What elements facilitated or hindered this experience?"*.

We provide some qualitative feedback from post session interviews.

Overall, participants express high satisfaction with the content and the practice of resilience skills. Several participants noted that the virtual practice was highly effective and educational for their needs, with some expressing a preference for this method over traditional face-to-face training and lectures. As one participant remarked: *"It's wow, much better than just sitting in a lecture, more comfortable, it feels like real practice. It serves as a space for learning and practicing because you can really see the content you are learning"* (female, social worker). Another participant shared: *"The learning experience with this method allowed me to make mistakes more freely and learn from them without the fear of being judged"* (female, resilience coach).

Some participants from the resilience program commented on the differences between the modules taught in their training and those in the virtual experience. They recommended incorporating the modules they use more consistently, along with the worksheets and guidelines they practice with, to better align the virtual simulation with their existing training. Additionally, some participants, particularly the more experienced ones, noted that the scenarios were sometimes too simple and straightforward: *"Sometimes it was too easy; I would have liked it to be a bit more complex, like in reality. Maybe there could be levels in terms of professionalism- it was a bit too basic"* (female, social worker). These remarks highlight the potential importance of the *Inceptor*, which allows tailoring VR training sessions to multiple types of users, various levels of expertise/difficulty, diverse backgrounds, and more.

Many participants expressed high satisfaction with the simulation, noting that it could significantly contribute to their confidence in actual scenarios with real patients. As one participant remarked: *"A lot (in terms of confidence). It's about practicing presence in the room, practicing saying things, being exposed to patients. The exposure in VR is no less real. It really contributes to practice, being one-on-one with a patient"* (male, psychiatric nurse). Another participant shared: *"It really helps with building confidence. I didn't think I would feel this way; I thought it would be less human. If Rachel had been a bit more human-like, it would have been perfect, and I would have felt 100%"* (female, resilience coach).

Some participants noted that the realism of the virtual characters could be improved, particularly regarding their movements, which were often described as too frequent (too jittery) and sometimes unnatural. As mentioned above, we have used an interim version of the content for the evaluation, in which the number of animation clips in the library was extremely restricted.

The analysis revealed a need for greater flexibility and adaptability within the simulation. Some participants expressed a desire for more freedom in their responses, rather than being limited to predefined options, suggesting that more flexible scenarios could enhance the learning experience. A few participants suggested allowing users to speak freely instead of reading pre-written text. As one participant mentioned: *"It's too structured, I'm not really learning. When I get a multiple-choice answer, it feels less meaningful to me. I wouldn't do it again unless it was more interactive. These concepts are familiar to me, but yes, it could help deepen my understanding, reinforce my skills, and explore different styles, especially if I could bring more of myself into it"* (male, resilience coach).

Participants also emphasized the need for varied scenarios that could be tailored to different patient types across various departments or sectors, as well as to different levels of experience and expertise: *"I think additional response options could be added to allow the trainee to practice with different types of patients who have varying levels of emotional awareness, and to adjust my responses accordingly"* (female, psychiatrist). Another participant noted: *"I felt that it simulates real-world interventions, but it doesn't simulate patients in psychiatric closed wards. There needs to be more variety and adaptation of the content for our department"* (female, psychiatric nurse).

5 Discussion

The paper introduced *Inceptor*, a story-first, human-in-the-loop engine that converts ordinary text scripts into production-ready social-VR scenes and demonstrated its viability through a 32-scenario resilience-training programme. Our deployment shows that domain experts with minimal technical background can generate hours of credible VR content in days rather than weeks, provided that key creative decisions—story structure, clinical phrasing, pedagogical pacing—remain under their direct control.

At the same time, the animation-selection study highlights a productive middle ground between manual craft and full automation: large-language-model prompting supplied approximately "novice-level" body-language choices that the expert could accept or override, cutting authoring effort without eroding authorial intent.

These findings suggest a future in which generative AI handles micro-level variation—prosody, gaze, gestural nuance—while authors retain ownership of macro-level narrative and ethical constraints. Realising that vision will require richer perceptual models to read user affect and tighter authoring controls to bound agent autonomy, but the present results indicate that such a hybrid approach can already deliver engaging, goal-directed social interactions at a fraction of traditional production cost.

5.1 Limitations

The *Inceptor* is far from a general purpose text to 3D conversion engine; it is useful for generating social scenarios. Moreover, it primarily focuses on verbal and non-verbal communication, requiring users to employ additional resources for generating avatars and scenes, and does not support animations involving interactions with objects,

moving or changing posture between standing and sitting, which limits its scope to communication-only scenarios. Despite this limitation, the use cases for *Inceptor* are important for specific training and counseling contexts. Furthermore, the architecture of the tool allows for future extensions and enhancements. This flexibility suggests that while *Inceptor* is not a comprehensive solution for all VR needs, it offers significant capabilities within its specialized domain, providing a foundation for expanded functionality in future developments.

5.2 Ethical Considerations

The design and deployment of virtual patient simulations for resilience training requires attention to several ethical considerations. First, participant data must be collected, stored, and processed in compliance with relevant privacy regulations (e.g., GDPR, HIPAA) to safeguard personal information and sensitive health data. Second, the immersive nature of VR may invoke strong emotional reactions, necessitating clear informed consent procedures and accessible support for participants who may experience distress. Additionally, the virtual scenarios should be validated for cultural sensitivity and inclusiveness to ensure equitable access and minimize unintended bias. Finally, ongoing collaboration with ethics committees and domain experts is essential to maintain transparent protocols and to continuously refine the system with participant well-being as a central focus.

5.3 Future Work

While our initial focus has been on linear scenarios, the *Inceptor* currently supports a limited model of interactivity. Future work could enhance this aspect, incorporating more sophisticated AI-driven dialogue systems, including LLM-based interactions (e.g., see [30]). There is a tradeoff between high quality human crafted experiences based on predefined narrative, vs real time simulation, resulting in emergent narrative. Further research is required in bridging this gap between our 'story-first' approach and applications based on intelligent autonomous agents with human-like behavior.

While a story-first design deliberately limits agent autonomy, selectively adding a degree of non-verbal awareness could improve moment-to-moment credibility without letting characters "wander off script". Recognizing simple behavioral cues from the participant—-a head-nod, a shrug, sustained eye-contact—-would allow a virtual counsellor to deliver the same scripted line with an appropriate prosody, gesture, or timing adjustment, thereby appearing more attentive while still advancing the predetermined plot. The research challenge is to find the sweet spot: enough perceptual intelligence for graceful, context-sensitive reactions, yet constrained sufficiently so the experience remains faithful to the authored narrative.

Additionally, there is scope for more nuanced animation generation, particularly in automating behaviors such as speaking and other non-verbal cues. It is possible to replace our approach, based on selecting animation clips from a pre-existing library, with generative models that can automatically generate animation to match spoken voice (e.g., [3, 10]). While our evaluations suggest that the current character models and animation meet basic requirements, further refinement could lead to more lifelike and believable characters, significantly enhancing user experience.

Our platform supports both VR and flat-screen web content, both of which can be generated automatically. In practice, this creates a trade-off between the immersive, high-impact nature of VR—an experience participants often find uniquely engaging—and the convenience of web-based access, which can be used anytime, anywhere. Although Unity-generated flat-screen content can be produced seamlessly, it may appear "cartoonish" to users who expect photorealism in a 2D environment. Achieving full photorealism in VR is extremely challenging, while doing so for 2D is more feasible but requires substantially different production pipelines. Consequently, further investigations are needed to determine whether the benefits of VR's enhanced immersion sufficiently outweigh the accessibility advantages of flat-screen content, particularly with regard to long-term user engagement and skill acquisition.

5.4 Conclusion

The evaluation study with the content generated by the *Inceptor* is encouraging. Participants expressed a desire for increased flexibility and diverse scenarios that could be adapted to different patient types, departments, and levels of expertise. These insights emphasize the significance of the *Inceptor* framework, which enables easy modifications and customization. Tailoring scenarios to specific requirements and professional levels could ultimately demonstrate the tool's effectiveness across a wider range of contexts.

We intend to release *Inceptor* as an open-source tool under the AFL 3.0 license, facilitating wide accessibility and community-driven enhancements. Our experience with *Inceptor* has demonstrated its significant utility in social science education and research in our lab, offering robust support for creating interactive and engaging scenarios that enhance learning and data collection. Moreover, the tool has been effectively used to develop applied projects that have been deployed in real-world settings. We encourage the community to leverage this tool to explore new possibilities in VR and social science research, pushing the boundaries of what can be achieved with open-source technology in these fields.

Acknowledgments

This work was partially supported by projects GuestXR (#101017884) and Socrates EU projects (#951930), which have received funding from the European Union's Horizon 2020 research and innovation program. The work was also partially supported by a philanthropic donation from Union Group, Israel.

The authors wish to thank Niv Rosnovsky for his involvement in programming the Inceptor, Elior Bengelsdorf for his support in the VR production of the resilience project, Tal Nakash and Prof. Anat Klomek Brunstein for the content of the resilience project, and Stav Schreiber for his overall support in managing the resilience project.

References

[1] American Psychological Association. n.d.. *Resilience*. https://www.apa.org/topics/resilience Accessed: 2025-01-18.

[2] Peter Ardhianto, Yonathan Purbo Santosa, Christian Moniaga, Maya Putri Utami, Christine Dewi, Henoch Juli Christanto, and Abbott Po Shun Chen. 2023. Generative deep learning for visual animation in landscapes design. *Scientific Programming* 2023, 1 (2023), 9443704.

[3] Eiichi Asakawa, Naoshi Kaneko, Dai Hasegawa, and Shinichi Shirakawa. 2022. Evaluation of text-to-gesture generation model using convolutional neural network. *Neural Networks* 151 (2022), 365–375. doi:10.1016/j.neunet.2022.03.041

[4] X Bai and Others. 2020. High-fidelity Facial Avatar Reconstruction from Monocular Video with Generative Priors. https://github.com/bbaaii/HFA-GP. *GitHub Repository* (2020).

[5] F Banfi, Stephen Fai, R Brumana, et al. 2017. BIM automation: advanced modeling generative process for complex structures. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. Copernicus GmbH, 9–16.

[6] Tessa Beinema, Daniel Davison, Dennis Reidsma, Oresti Banos, Merijn Bruijnes, Brice Donval, Álvaro Fides Valero, Dirk Heylen, Dennis Hofs, Gerwin Huizing, et al. 2021. Agents United: An open platform for multi-agent conversational systems. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*. 17–24.

[7] Mengyu Chen, Marko Peljhan, and Misha Sra. 2024. ConnectVR: A Trigger-Action Interface for Creating Agent-based Interactive VR Stories. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. IEEE, 286–297.

[8] T. D. Cosco, K. Howse, and C. Brayne. 2017. Healthy ageing, resilience and wellbeing. *Epidemiology and Psychiatric Sciences* 26, 6 (2017), 579–583. doi:10.1017/S2045796017000324

[9] Jacob Garbe, Max Kreminski, Ben Samuel, Noah Wardrip-Fruin, and Michael Mateas. 2019. StoryAssembler: an engine for generating dynamic choice-driven narratives. In *Proceedings of the 14th International Conference on the Foundations of Digital Games*. 1–10.

[10] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. 2019. Learning individual styles of conversational gesture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3497–3506.

[11] Daniele Giunchi, Nels Numan, Elia Gatti, and Anthony Steed. 2024. DreamCodeVR: Towards Democratizing Behavior Design in Virtual Reality with Speech-Driven Programming. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. IEEE, 579–589.

[12] Chaoqun Gong, Yuqin Dai, Ronghui Li, Achun Bao, Jun Li, Jian Yang, Yachao Zhang, and Xiu Li. 2024. Text2Avatar: Text to 3d Human Avatar Generation with Codebook-Driven Body Controllable Attribute. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 16–20.

[13] Charlie Hargood, Mark J. Weal, and David E. Millard. 2018. The StoryPlaces Platform: Building a Web-Based Locative Hypertext System. In *Proceedings of the 29th on Hypertext and Social Media* (Baltimore, MD, USA) *(HT '18)*. Association for Computing Machinery, New York, NY, USA, 128–135. doi:10.1145/3209542.3209559

[14] Yuze He, Yushi Bai, Matthieu Lin, Jenny Sheng, Yubin Hu, Qi Wang, Yu-Hui Wen, and Yong-Jin Liu. 2024. Text-image conditioned diffusion for consistent text-to-3D generation. *Computer Aided Geometric Design* 111 (2024), 102292.

[15] Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Bazavan, Mihai Fieraru, and Cristian Sminchisescu. 2024. Dreamhuman: Animatable 3d avatars from text. *Advances in Neural Information Processing Systems* 36 (2024).

[16] Canlin Li, Chao Yin, Jiajie Lu, and Lizhuang Ma. 2009. Automatic 3D scene generation based on Maya. In *2009 IEEE 10th International Conference on Computer-Aided Industrial Design & Conceptual Design*. IEEE, 981–985.

[17] Chris Martens and Owais Iqbal. 2019. Villanelle: An authoring tool for autonomous characters in interactive fiction. In *Interactive Storytelling: 12th International Conference on Interactive Digital Storytelling, ICIDS 2019, Little Cottonwood Canyon, UT, USA, November 19–22, 2019, Proceedings 12*. Springer, 290–303.

[18] Joshua McCoy, Mike Treanor, Ben Samuel, Aaron A. Reed, Michael Mateas, and Noah Wardrip-Fruin. 2014. Social Story Worlds With Comme il Faut. *IEEE Transactions on Computational Intelligence and AI in Games* 6, 2 (2014), 97–112. doi:10.1109/TCIAIG.2014.2304692

[19] Tom Monnier, Jake Austin, Angjoo Kanazawa, Alexei Efros, and Mathieu Aubry. 2023. Differentiable blocks world: Qualitative 3d decomposition by rendering primitives. *Advances in Neural Information Processing Systems* 36 (2023), 5791–5807.

[20] Charlie Nash and Christopher KI Williams. 2017. The shape variational autoencoder: A deep generative model of part-segmented 3d objects. In *Computer Graphics Forum*, Vol. 36. Wiley Online Library, 1–12.

[21] Kyungjin Park, Bradford W. Mott, Wookhee Min, Kristy Elizabeth Boyer, Eric N. Wiebe, and James C. Lester. 2019. Generating Educational Game Levels with Multistep Deep Convolutional Generative Adversarial Networks. In *2019 IEEE Conference on Games (CoG)*. 1–8. doi:10.1109/CIG.2019.8848085

[22] Dan Pollak, Jonathan Giron, and Doron Friedman. 2023. Inceptor: an open source tool for automated creation of 3D social scenarios. In *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 474–476.

[23] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988* (2022).

[24] Jane C Richardson and Carolyn A Chew-Graham. 2016. Resilience and well-being. *Mental Health and Older people: a guide for primary care practitioners* (2016), 9–17.

[25] Mark O Riedl and Robert Michael Young. 2010. Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research* 39 (2010), 217–268.

[26] James Ryan, Ethan Seither, Michael Mateas, and Noah Wardrip-Fruin. 2016. Expressionist: An authoring tool for in-game text generation. In *Interactive Storytelling: 9th International Conference on Interactive Digital Storytelling, ICIDS 2016, Los Angeles, CA, USA, November 15–18, 2016, Proceedings 9*. Springer, 221–233.

[27] Marie-Laure Ryan. 2009. From narrative games to playable stories: Toward a poetics of interactive narrative. *StoryWorlds: a journal of narrative studies* 1 (2009), 43–59.

[28] Ben Samuel, Aaron A Reed, Paul Maddaloni, Michael Mateas, and Noah Wardrip-Fruin. 2015. The ensemble engine: Next-generation social physics. In *Proceedings of the Tenth International Conference on the Foundations of Digital Games (FDG 2015)*. 22–25.

[29] Leixian Shen, Haotian Li, Yun Wang, and Huamin Qu. 2024. From Data to Story: Towards Automatic Animated Data Video Creation with LLM-based Multi-Agent Systems. *arXiv preprint arXiv:2408.03876* (2024).

[30] Alon Shoa, Ramon Oliva, Mel Slater, and Doron Friedman. 2023. Sushi with Einstein: Enhancing Hybrid Live Events with LLM-Based Virtual Humans. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*. 1–6.

[31] Bryan Temprado-Battad, José-Luis Sierra, and Antonio Sarasa-Cabezuelo. 2019. An online authoring tool for interactive fiction. In *2019 23rd International Conference Information Visualisation (IV)*. IEEE, 339–344.

[32] Christina Tsalicoglou, Fabian Manhardt, Alessio Tonioni, Michael Niemeyer, and Federico Tombari. 2024. Textmesh: Generation of realistic 3d meshes from text prompts. In *2024 International Conference on 3D Vision (3DV)*. IEEE, 1554–1563.

[33] Weiyue Wang, Qiangui Huang, Suya You, Chao Yang, and Ulrich Neumann. 2017. Shape inpainting using 3d generative adversarial network and recurrent convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 2298–2306.

[34] Xiao Yang, Chang Liu, Longlong Xu, Yikai Wang, Yinpeng Dong, Ning Chen, Hang Su, and Jun Zhu. 2023. Towards effective adversarial textured 3d meshes on physical face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4119–4128.

[35] Sheng Ye, Yubin Hu, Matthieu Lin, Yu-Hui Wen, Wang Zhao, Yong-Jin Liu, and Wenping Wang. 2024. Indoor Scene Reconstruction with Fine-Grained Details Using Hybrid Representation and Normal Prior Enhancement. *IEEE Transactions on Visualization and Computer Graphics* (2024).

[36] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. 2024. CLAY: A Controllable Large-scale Generative Model for Creating High-quality 3D Assets. *ACM Transactions on Graphics (TOG)* 43, 4 (2024), 1–20.