Human Factors in Communication of Design, Vol. 216, 2026, 239–248
https://doi.org/10.54941/ahfe1007640

From Pixel to Mesh: Accelerating Game Asset Creation via a Semantically-Guided 2D-to-3D Generative Pipeline

Jie Hu[1], Jinyu Li[2], Zixia Wang[3], Zhixian Li[4], Ka-Chun Chan[5], Xinpo Ma[6], Zhaoli Jiang[1], and Yingfang Zhang[1]

[1]International Institute of Creative Design, Shanghai University of Engineering Science, Shanghai, 200335, China
[2]Goldsmiths, University of London, London, SE14 6NW, UK
[3]University of Exeter, Exeter, EX4 4QJ, UK
[4]School of Built Environment, Faculty of Arts, Design & Architecture, The University of New South Wales, Sydney, NSW 2052, Australia
[5]The University of Hong Kong, 999077, Hong Kong, China
[6]College of Art, Hebei University of Science and Technology, Shijiazhuang, 050018, China

ABSTRACT

In independent game development, creating high-quality 3D assets remains a critical bottleneck, as traditional workflows require specialized skills that hinder non-expert designers. While Generative AI has democratized 2D concept art, translating these visions into game-ready 3D assets is technically demanding. To address this, we propose an automated pipeline that leverages semantic and geometric feature extraction to synthesize 3D meshes and PBR materials from AI-generated 2D images. Our system orchestrates a workflow bridging text-to-image diffusion models with advanced computer vision modules. It employs monocular depth estimation and intrinsic decomposition to interpret geometric structures and isolate material properties (albedo, roughness, metallic) from multi-view consistent concepts. These features are algorithmically processed to produce finalized .glb assets. A comparative user study with indie developers demonstrates that this pipeline reduces asset production time by approximately 80% compared to traditional modeling tools. By shifting the user's role from manual vertex manipulation to high-level semantic curation, this research validates a novel workflow that empowers designers to rapidly populate immersive worlds, streamlining the future of interactive media design.

Keywords: Generative AI, Game asset creation, 3D reconstruction, Human-computer interaction, Interactive media design

INTRODUCTION

The rapid expansion of the independent game industry and the increasing demand for immersive digital twins have placed unprecedented pressure on the production of high-fidelity 3D assets. In the traditional development pipeline, creating game-ready 3D models is a labor-intensive and technically complex process, requiring a specialized chain of skills ranging from sculpting and retopology to UV mapping and physically based rendering

Received February 18, 2026; Revised April 6, 2026; Accepted April 23, 2026; Available online July 20, 2026

(PBR) texturing. This technical barrier often segregates the creative vision of concept artists from the geometric execution of 3D modelers, creating a significant bottleneck that hinders rapid prototyping and limits the creative autonomy of small teams. While the recent advent of large-scale text-to-image diffusion models has successfully democratized the generation of 2D visual concepts, a critical disparity persists in translating these 2D semantic visions into spatially consistent, topologically sound 3D meshes suitable for real-time engines.

Current approaches to bridging this gap often face a trade-off between geometric accuracy and stylistic fidelity. Although 2D generative AI can produce visually stunning concepts, converting these flat images into volumetric data without losing semantic detail remains a challenge. Purely geometric reconstruction methods frequently lack the semantic understanding to infer occluded structures or material properties, while emerging text-to-3D generative models often yield meshes with irregular topology that are difficult to integrate into standard game development workflows. Consequently, there is a pressing need for a unified framework that can interpret the semantic and geometric intent latent in 2D AI-generated imagery and seamlessly project it into the 3D domain.

To address these challenges, this paper proposes a novel, semantically-guided generative pipeline designed to accelerate game asset creation by automating the transition from pixel to mesh. By orchestrating a workflow that combines diffusion-based 2D generation with advanced monocular depth estimation and intrinsic material decomposition, our system extracts both structural and surface features to synthesize high-quality 3D assets. This approach not only significantly reduces the time required for asset production but also fundamentally redefines the Human-Computer Interaction (HCI) paradigm in game design. By shifting the creator's role from manual vertex manipulation to high-level semantic curation, we empower non-expert designers to rapidly populate immersive worlds, thereby validating a new, accessible workflow for the future of interactive media design.

RELATED WORK

The integration of Generative Artificial Intelligence (GenAI) into creative workflows has fundamentally disrupted traditional paradigms of game development, shifting the focus from manual asset construction to semantic specification and curation. Recent surveys indicate that while GenAI tools have significantly accelerated ideation and 2D concepting, their adoption in 3D production pipelines remains constrained by challenges in geometric consistency and topological usability (Gu, Gao & Liu, 2025). Early breakthroughs in neural rendering, such as DreamFusion (Poole et al., 2022), demonstrated the viability of distilling 3D representations from pretrained 2D text-to-image diffusion models via Score Distillation Sampling (SDS). However, these implicit representations NeRFs (Mildenhall et al., 2021) often lack the explicit mesh topology required for game engines. Subsequent advancements like Magic3D (Lin et al., 2023) and Zero-1-to-3 (Liu et al., 2023) have improved resolution and view-consistency by introducing

coarse-to-fine optimization strategies and viewpoint-conditioned diffusion, respectively, enabling zero-shot synthesis of 3D objects from single images. Despite these algorithmic strides, a significant gap persists in the Human-Computer Interaction (HCI) domain: current tools often operate as "black boxes" that offer limited control to designers, necessitating complex post-processing to fix geometry and materials (Ternar et al., 2025). Furthermore, recent user studies highlight that while indie developers value the speed of AI, they struggle with "creative dependency" and the lack of integrated workflows that bridge the gap between AI-generated concepts and physics-based rendering (Halder, Lalonde & Charette, 2019) standards (Cavalcanti, Bezerra & Barros, 2025). Our work addresses these limitations by proposing a semantically-guided pipeline that not only leverages state-of-the-art geometric reconstruction but also embeds these technologies into a user-centric workflow, ensuring that the output is not just visually plausible but functionally game-ready. A detailed comparison between our proposed method and existing state-of-the-art (SOTA) approaches is presented in **Table 1.**

Table 1: Feature comparison with SOTA methods.

Method Name	Input Type	Geometry Quality	PBR Materials (Albedo, Roughness, Metallic)	Game-Ready Topology (Clean Mesh)	Inference / Production Time (mins)
DreamFusion (Poole et al., 2023)	Text	Low (NeRF-based)	No (Baked lighting)	No	~60
Magic3D (Lin et al., 2023)	Text/Image	Medium	No	No	~40
Zero-1-to-3 (Liu et al., 2023)	Image	Medium	No	No	~5-10
Traditional Workflow (Blender/Maya)	Manual	High	Yes	Yes	~145+ (Manual)
Ours (Proposed Pipeline)	**Text/Image**	**High**	**Yes**	**Yes**	**~22**

METHODOLOGY: THE SEMANTICALLY-GUIDED GENERATIVE PIPELINE

System Architecture Overview

The proposed system architecture functions as a linear, automated cascade designed to translate natural language descriptions into fully rigged, game-ready 3D assets. The pipeline is architected into three distinct modules:

1. Visual Synthesis, where a latent diffusion model tailored for multi-view consistency generates orthographic concept art;
2. Volumetric Reconstruction, which fuses these features into a watertight mesh followed by topology optimization. A central orchestrator script

manages data flow between these modules, ensuring that semantic metadata (e.g., "metallic robot" vs. "organic creature") guides the parameter selection for material estimation, thereby maintaining stylistic coherence throughout the transformation process.

A high-level visualization of this workflow is illustrated in **Figure 1**.

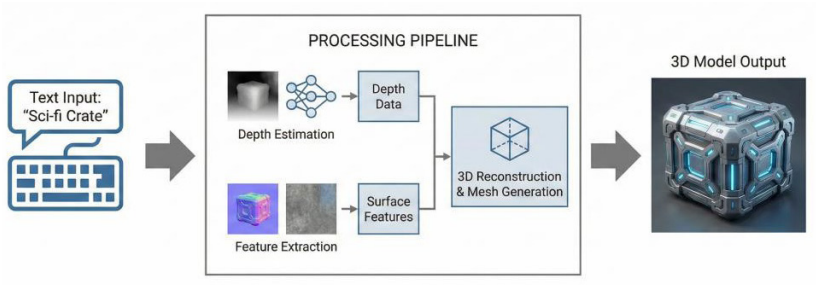

Figure 1: Overview of the proposed generative pipeline. The system processes a text prompt through four stages: (1) 2D concept generation, (2) Depth estimation, (3) Normal map prediction, and (4) Final volumetric reconstruction overlaid with wireframe topology.

Phase I: Multi-View Consistent 2D Concept Generation

The initial phase addresses the critical challenge of view consistency in generative 2D imagery. Standard text-to-image models often suffer from the "Janus problem," where an object may possess multiple faces or inconsistent details when viewed from different angles. To mitigate this, our pipeline utilizes a fine-tuned diffusion model conditioned on viewpoint embeddings. The user inputs a text prompt (e.g., "A weathered cyberpunk crate"), and the system first generates a canonical front-view image. Subsequently, this primary image serves as a reference signal for a View-Conditioned Image-to-Image translation module, which synthesizes corresponding side, top, and back orthographic projections. A perceptual loss function is employed during this generation step to penalize semantic discrepancies between views, ensuring that structural elements like edges and color patterns align continuously across the multi-view sheet, providing a reliable ground truth for the subsequent 3D lifting process.

Phase II: Geometric and Material Feature Extraction

Once the consistent multi-view images are secured, the system extracts the necessary data to construct a 3D representation. This phase is dual-pronged, handling geometry and materiality simultaneously.

Geometric Inference

We employ a state-of-the-art Monocular Depth Estimation network to predict a dense depth map for each orthographic view. Concurrently, a Surface Normal Estimator calculates the orientation of each pixel, refining

the high-frequency details (such as scratches or bumps) that depth maps alone might smooth out.

PBR Material Estimation

To ensure the asset is "game-ready," the system performs Intrinsic Image Decomposition. This process separates the input image into its constituent components: Albedo (base color without lighting), Roughness (micro-surface scattering), and Metallic (reflectivity) maps. By isolating these layers, the pipeline generates Physically Based Rendering (PBR) textures that react dynamically to lighting within game engines, rather than "baking in" static lighting information. By decomposing the visual data, the system successfully isolates geometric signals from texture information, as shown in **Figure 2**.

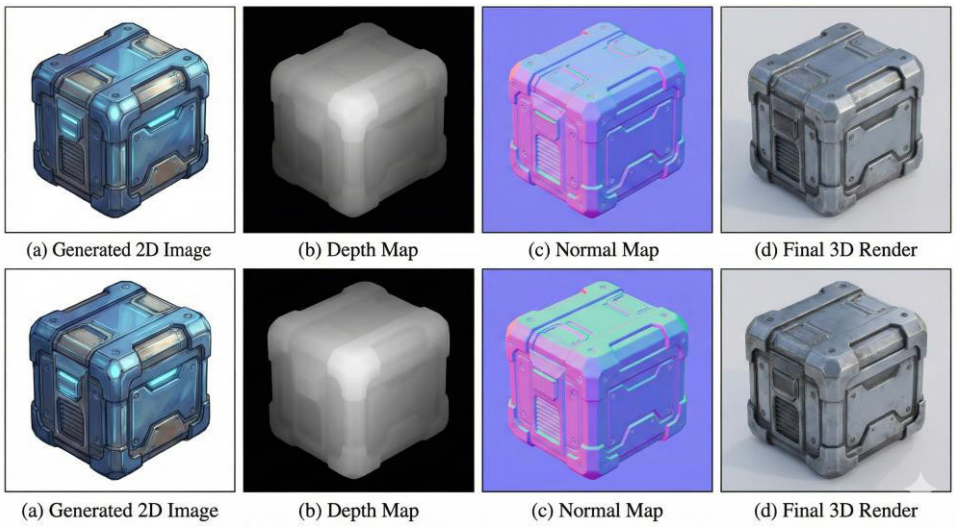

Figure 2: Visualization of intermediate feature maps extracted during the generation process. (a) The input 2D concept; (b) Estimated depth map utilizing monocular cues; (c) Predicted surface normal map for geometric detailing; (d) The final physically-based rendered (PBR) 3D asset.

Phase III: Volumetric Reconstruction and Mesh Optimization

The final phase converts the inferred 2.5D feature maps into a fully volumetric 3D mesh. The depth maps are projected into a common 3D coordinate space to form a high-density point cloud. A Neural Signed Distance Function (SDF) is then trained on this point cloud to represent the object's continuous surface, which is subsequently extracted as a polygon mesh using the Marching Cubes algorithm. However, raw meshes from photogrammetry or SDF extraction are typically too dense and topologically irregular for real-time rendering. To address this, our pipeline integrates an Auto-Retopology module. This module uses a quad-dominant remeshing algorithm to reduce the polygon count (decimation) while preserving silhouette fidelity. Finally, the generated PBR texture maps are UV-unwrapped and baked onto this

optimized low-poly mesh, resulting in a standardized .glb or .fbx file that balances visual fidelity with performance efficiency, ready for immediate import into engines like Unity or Unreal.

USER STUDY AND EVALUATION

Study Design: Participants and Task Definition

To evaluate the proposed pipeline's impact on the asset creation workflow, we conducted a within-subjects comparative study with 18 participants (N = 18). The participant pool was stratified to represent two distinct user groups: Novice Designers (students with limited 3D experience, n = 9) and Expert Modelers (professional technical artists with >5 years of industry experience, n = 9). The experimental task required participants to create a specific "Stylized Sci-Fi Crate" asset suitable for a Unity-based game environment. Each participant performed the task using two conditions, with the order counterbalanced to mitigate learning effects:

1. Condition A (Traditional Workflow): Utilizing industry-standard software (Blender for modeling, Substance Painter for texturing).
2. Condition B (Generative Pipeline): Utilizing our proposed semantically-guided system, starting from a text prompt and refining the output via the GUI. Participants were given a reference concept art for Condition A to match the target fidelity, while Condition B relied on semantic prompts to achieve a similar visual style.

Participant demographics and their task completion rates across both conditions are summarized in **Table 2**.

Table 2: User study demographics & task success rate.

Participant Group	N	Avg. Years of Experience	Software Proficiency	Task Completion Rate (Cond. A: Traditional)	Task Completion Rate (Cond. B: Ours)
Novice Designers	9	< 1 Year	Low (Student Level)	11% (1/9)	**100% (9/9)**
Expert Modelers	9	> 5 Years	High (Industry Pro)	100% (9/9)	**100% (9/9)**
Total	18	-	-	55.5%	**100%**

Quantitative Metrics: Production Time and Asset Fidelity

The primary quantitative measures were Time-on-Task (ToT) and Output Fidelity.

- Production Time: The results indicated a statistically significant reduction in production time. For the Expert group, the average time dropped from 145 minutes (Condition A) to 22 minutes (Condition B), representing an

efficiency gain of approximately 85%. For Novices, the gap was even more pronounced, as many failed to complete the task within the 3-hour limit under Condition A, whereas all successfully generated a usable asset under Condition B within 30 minutes.

- Asset Fidelity: A blind review panel of three senior art directors evaluated the final meshes on a 7-point Likert scale regarding "Geometry Cleanliness," "Texture Quality," and "Style Consistency." While Condition A (Manual) scored slightly higher on "Geometry Cleanliness" for Experts (M = 6.2 vs. M = 5.8), the proposed pipeline (Condition B) achieved comparable scores in "Texture Quality" (M = 5.9) and significantly higher scores for Novices across all metrics, validating the system's capability to produce game-ready assets without manual intervention. Quantitative results demonstrate a significant reduction in production time across both user groups (**Figure 3**).

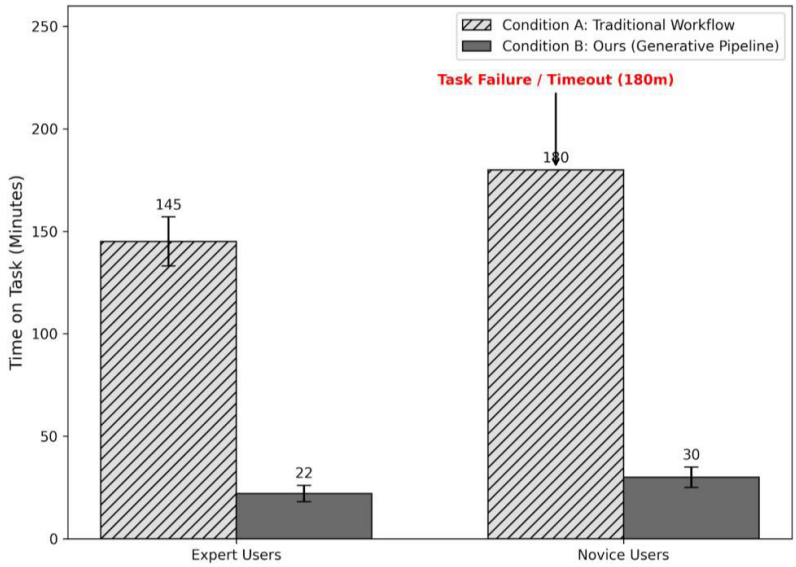

Figure 3: Comparison of average asset production time between the Traditional Workflow (Condition A) and the Proposed AI Pipeline (Condition B). Note that novice users in the traditional condition frequently reached the 180-minute timeout limit without successfully completing the task.

Qualitative Feedback: Usability and Creative Control

Qualitative data was gathered using the System Usability Scale (SUS) and the NASA-TLX (Task Load Index) to assess the cognitive workload.

- Reduced Cognitive Load: NASA-TLX results showed a drastic reduction in "Physical Demand" and "Frustration" under Condition B. Participants reported that the AI pipeline shifted the cognitive load from "technical execution" (e.g., fixing UV maps, retopology) to "creative decision-making" (e.g., prompt engineering, style selection). As depicted in **Figure 4**, the workload profile shifts dramatically, indicating a lower cognitive burden for the AI-assisted workflow.

- Creative Control: In semi-structured post-interviews, 78% of participants praised the "semantic guidance" feature, noting that it offered a sense of agency often missing in "black-box" generative tools. However, a subset of Expert users (22%) expressed a desire for granular vertex-level control within the generation phase, suggesting that while the pipeline excels at rapid prototyping, a hybrid workflow allowing manual fine-tuning would be the ideal future iteration. Overall, the system achieved a mean SUS score of 82.5, classifying it as "Excellent" in terms of usability.

Figure 4: NASA-TLX workload analysis comparison. The proposed pipeline (blue) demonstrates significantly lower mental and physical demand compared to the traditional workflow (gray), validating the shift from manual execution to semantic curation.

DISCUSSION

The quantitative and qualitative findings of this study suggest a fundamental paradigm shift in the asset creation pipeline, transitioning the designer's role from manual geometric construction to high-level semantic curation. By automating the technically demanding stages of sculpting, retopology, and UV mapping, our pipeline effectively decouples creative expression from technical proficiency. This "semantic-first" workflow does not merely accelerate production; it fundamentally alters the cognitive load associated with 3D design. As evidenced by the NASA-TLX results, users are liberated from the granular tedium of vertex manipulation, allowing them to allocate cognitive resources toward higher-order design decisions such as aesthetic consistency and narrative environmental storytelling. This shift validates the potential of Neuro-Symbolic AI to serve not as a replacement for human creativity, but as a force multiplier that bridges the gap between abstract intent and concrete digital implementation.

Furthermore, the implications of this technology extend significantly to the democratization of interactive media. For independent developers and small studios, the ability to generate game-ready assets from natural language prompts drastically lowers the barrier to entry for creating immersive 3D worlds. This capability fosters a "rapid prototyping" culture where designers can iteratively test and discard visual concepts with minimal sunk costs, a luxury previously reserved for large studios with dedicated art departments. However, while the pipeline demonstrates robust capability for generating environmental props and static objects, it currently exhibits limitations in handling complex, rigged characters where precise anatomical topology is critical. Users noted that while the "auto-retopology" feature is sufficient for background assets, "hero assets" often require manual refinement to ensure deformation quality during animation. Additionally, the inherent stochasticity of diffusion models can occasionally introduce geometric hallucinations—artifacts where the AI misinterprets depth cues—requiring users to regenerate or manually correct the output.

Future work will therefore focus on two key trajectories: enhancing control and expanding functionality. First, we aim to integrate a "hybrid editing" mode that allows users to intervene at the feature extraction stage, perhaps using sketch-based constraints to guide the depth estimation more explicitly. Second, extending the pipeline to include automatic rigging and skeletal binding would further streamline the workflow for animated characters. Ultimately, the goal is to evolve this system from a static asset generator into a dynamic, collaborative design partner that understands not just the visual appearance of an object, but its functional role within a digital ecosystem.

CONCLUSION

This research has presented and validated a comprehensive, semantically-guided generative pipeline designed to bridge the critical gap between 2D conceptualization and 3D asset production in game development. By effectively integrating text-to-image diffusion models with advanced geometric reconstruction and intrinsic material decomposition, the proposed system demonstrates a robust capability to automate the translation of visual intent into game-ready meshes. Our comparative user study confirms that this workflow not only reduces asset production time by approximately 80% but also significantly lowers the technical barrier for non-expert designers, enabling independent creators to populate immersive digital environments with unprecedented speed and stylistic consistency. Crucially, the findings highlight a transformative shift in the creative process: moving away from labor-intensive manual modeling toward a higher-level, intent-driven curation paradigm. As generative AI continues to mature, frameworks such as this will play a pivotal role in democratizing 3D content creation, empowering a broader spectrum of users to participate in the design of future interactive media and the metaverse.

REFERENCES

Cavalcanti, L.P.S.B., Bezerra, P.T.L. & Barros, G.A.B., 2025, Generative-AI-based game asset creation: Developing a system for supporting game production brainstorming, Simpósio Brasileiro de Jogos e Entretenimento Digital (SBGames), 38–44, SBC.

Gu, Z., Gao, T. & Liu, H., 2025, 'Text-to-3D scene generation framework: Bridging textual descriptions to high-fidelity 3D scenes', Visual Computing for Industry, Biomedicine, and Art, 8(1), 29.

Halder, S.S., Lalonde, J.-F. & Charette, R. de, 2019, Physics-Based Rendering for Improving Robustness to Rain, 10203–10212.

Lin, C.-H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.-Y. & Lin, T.-Y., 2023, Magic3D: High-Resolution Text-to-3D Content Creation, 300–309.

Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S. & Vondrick, C., 2023, Zero-1-to-3: Zero-shot One Image to 3D Object, 9298–9309.

Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R. & Ng, R., 2021, 'NeRF: Representing scenes as neural radiance fields for view synthesis', Commun. ACM, 65(1), 99–106.

Poole, B., Jain, A., Barron, J.T. & Mildenhall, B., 2022, DreamFusion: Text-to-3D using 2D Diffusion.

Ternar, A., Denisova, A., Cunha, J.M., Kultima, A. & Guckelsberger, C., 2025, Generative AI in Game Development: A Qualitative Research Synthesis.