SpeLL: A Generative Framework for Synthesizing 3D Visual Effects from Text Prompts with Large Language Models

Jeongseok Oh
jungseok17@gm.gist.ac.kr
GIST
HCIS Lab
Gwangju, Republic of Korea

SeungJun Kim*
seungjun@gist.ac.kr
GIST
HCIS Lab
Gwangju, Republic of Korea



Figure 1: Example 3D VFX generated by SpeLL. SpeLL is a generative framework designed for professional VFX designers creating effects for 3D interactive content. The system synthesizes parameterized VFX assets from natural language prompts using LLMs, enabling parameter-based editing.

Abstract

Creating visual effects (VFX) for 3D content requires specialized design knowledge and technical expertise. Existing methods focus on creating VFX for videos or exploring designs for novices, while professional workflows need frameworks producing complete parameterized 3D assets. This study presents SpeLL, a generative framework that produces editable 3D VFX drafts from text prompts using large language models and retrieval-based asset assembly. A formative study with professional VFX designers identified three design requirements: preference for simplified prompting, need for VFX-specialized assets, and manual spatial-temporal coordination. SpeLL addresses these through prompt enhancement, a spatial design module for component retrieval and layout arrangement, and a temporal design module for activation sequencing. A user study with twenty professional designers evaluated SpeLL against ablated frameworks, 3D generation baselines, and manual workflows. Results showed SpeLL scored higher on intention alignment, spatial-temporal layout quality, and editability, and reduced the effort needed to reach editable VFX structures. Participants valued parameterized outputs over visual complexity, emphasizing workflow integration and manual refinement. These findings indicate that combining LLM reasoning with retrieval-based asset assembly can support early exploration and drafting in professional VFX workflows, while leaving final expressive control to designers.

CCS Concepts

• **Human-centered computing** → **Systems and tools for interaction design**; • **Computing methodologies** → **Artificial intelligence**.

Keywords

creativity support tool, graphics design, artificial intelligence

ACM Reference Format:
Jeongseok Oh and SeungJun Kim. 2026. SpeLL: A Generative Framework for Synthesizing 3D Visual Effects from Text Prompts with Large Language Models. In *The 39th Annual ACM Symposium on User Interface Software and Technology (UIST '26), November 02–05, 2026, Detroit, MI, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3830398.3830697

*Corresponding author.

Please use nonacm option or ACM Engage class to enable CC licenses
This work is licensed under a Creative Commons Attribution 4.0 International License.
UIST '26, November 02–05, 2026, Detroit, MI, USA
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2856-3/2026/11
https://doi.org/10.1145/3830398.3830697

1 INTRODUCTION

Visual effects (VFX) provide feedback, atmosphere, and narrative support in 3D games and interactive media [19, 42]. 3D VFX elements include effects such as fire and smoke, magical spell effects, and particle systems that respond to player actions [43, 50, 68]. During gameplay, VFX communicate state changes, hazards, and affordances while guiding player attention [39, 47, 50]. Beyond functional roles, VFX shape the visual style of games: through manipulation of color, contrast, and brightness they establish mood and atmosphere while maintaining visual consistency across environments [1, 12, 42, 47].

In interactive 3D content, VFX assets are represented as complex compositions of multiple components including meshes, textures, and particles [45, 55]. While text-to-3D approaches have shown progress in generating individual meshes and textures [22, 36, 53], these models require additional adaptation since they are not specifically designed for VFX generation. LLM-based approaches have recently emerged for VFX creation, proposing methods for physically realistic video effects from natural language instructions [26, 31] and kinematic-driven particle systems for novice users [68]. Although these methods address video rendering or design exploration, a unified framework that incorporates meshes, textures, and particles needs to be explored since professional workflows require editable VFX structures that can be refined after generation.

Large Language Models (LLMs) have demonstrated spatial and temporal reasoning capabilities that provide coarse priors for VFX coordination. In the spatial domain, prior work shows scene-level reasoning, object localization, and geometric constraint solving across multi-object scenes [10, 27, 28]. In the temporal domain, LLMs can track event progressions, infer causal relationships, and retain salient dependencies across long sequences [5, 17].

Building on these capabilities, we present SpeLL, a generative framework that creates editable VFX drafts with meshes, textures, and particles. A formative study with professional VFX designers revealed a preference for high-level prompting followed by manual control, leading us to design SpeLL as an AI-assisted drafting framework in which LLMs make preliminary compositional decisions: selecting components, determining spatial arrangements, and sequencing temporal relationships. Figure 1 shows examples generated by SpeLL, which yields parameterized VFX assets that designers can further adjust in Unity Engine. The main contributions of this work are:

- **Understanding Professional VFX Design Requirements**: Through formative interviews with 10 professional VFX designers, we identify key aspects of VFX creation workflows and establish design goals for AI-assisted VFX generation.
- **VFX Generation Framework**: We introduce SpeLL, an LLM-based generative framework that creates parameterized, editable VFX drafts from natural language descriptions.
- **Integrated Multi-Component VFX Design**: Based on our formative study findings, we develop a framework that considers meshes, textures, and particle systems as unified components for VFX asset creation.
- **Empirical Validation with Professional Designers**: Our user study with 20 professional VFX designers shows that SpeLL improves output quality, editability, and workflow efficiency compared with alternative approaches.

2 RELATED WORKS

2.1 LLM for Spatial and Temporal Reasoning

Recent work has established that LLMs possess spatial and temporal reasoning capabilities applicable to 3D content tasks. In the spatial domain, SceneGPT enables spatial reasoning and object localization through scene graphs without specialized 3D training [9], Scene-LLM unifies egocentric and scene-level information for 3D visual question answering [20], and VideoRefer Suite demonstrates fine-grained spatial-temporal object understanding [65]. In the temporal domain, LLMs can infer causal and temporal relations from annotated event pairs [29], order events within interactive narratives [32], and reason over extended sequences via spatio-temporal scene graphs [44].

However, these capabilities have been primarily applied to understanding tasks such as question answering and scene comprehension. Leveraging them for generative coordination, jointly arranging multiple components in space and time, remains underexplored. SpeLL applies these reasoning capabilities to VFX generation, using LLMs to coordinate component selection, spatial arrangement, and temporal sequencing, while addressing their limitations through retrieval-based refinement.

2.2 3D Content Generation and Authoring Tools

3D mesh generation has progressed from CNN-based reconstruction [59] and physics-aware VAE synthesis [58] to text-driven approaches [56], with diffusion models now central to state-of-the-art 3D asset generation [48, 60]. However, these models are designed to produce geometrically detailed, visually complex meshes and textures suited for general 3D objects. VFX assets have fundamentally different requirements: meshes serve as simple topology carriers for shader-driven effects, and textures are tileable noise patterns rather than representational imagery. As a result, the outputs of existing 3D generation models are not directly compatible with VFX production pipelines, which require parameterized components that support two-dimensional UV texture coordinates that map textures onto mesh surfaces, texture scrolling, and real-time parameter adjustment.

On the authoring side, professional workflows still require manual control and iterative refinement [13, 41]. Authoring tools address this by preserving designer control while reducing technical implementation effort [2]: VRCopilot provides modifiable AI-suggested room layouts [67], Chen et al. proposed text-guided mesh detailing that preserves designer control [13], and Style2Fab separates functional from aesthetic parts for independent modification [18]. These tools demonstrate the value of AI-generated starting points that designers can refine [42], but they focus on individual object authoring rather than the multi-component spatial-temporal coordination that VFX creation demands. SpeLL builds on this principle by extending AI-assisted authoring from single-object refinement to the coordinated assembly of meshes, textures, and particles across space and time.

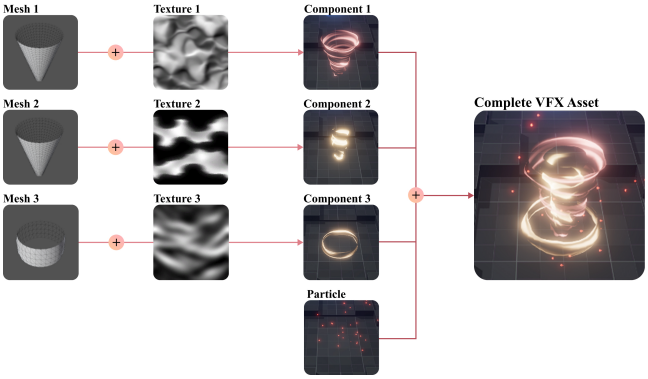


Figure 2: VFX assets are composed of simplified meshes combined with tileable noise textures.

2.3 VFX Synthesis with Text Prompts

Text-to-video diffusion models have introduced methods to directly synthesize VFX imagery from textual descriptions. MagicVFX [24] uses pre-trained video diffusion models to synthesize visual effects, constructing VFX-307, a dataset containing 675 videos across 15 effect categories. VFX Creator [37] presents a controllable VFX generation framework based on Video Diffusion Transformers. Google's Genie generates entire interactive 2D game environments from text prompts through spatiotemporal transformers instead of creating individual assets [8]. However, these approaches remain in the 2D domain and are not suitable for 3D content development.

Within 3D game engines, particle systems remain the primary VFX authoring workflow [25, 38], and prior work has explored particle-based effect construction [4, 16, 61, 66]. Recent LLM-based approaches have extended this direction: KinemaFX provides kinematic driven particle exploration for novice users [68], and AutoVFX decomposes VFX creation into executable sub-tasks via code generation [26]. However, these approaches focus on particle systems alone or target video-level rendering, leaving the integrated generation of meshes, textures, and particles for professional 3D workflows relatively underexplored. SpeLL addresses this gap by generating complete, parameterized VFX assets that combine all three component types within Unity Engine, supporting the parameter-based refinement workflows that professionals rely on.

3 FORMATIVE STUDY

To understand professional perspectives on VFX creation workflows and identify directions for AI assistance, we conducted formative semi-structured interviews with professional VFX designers.

3.1 Formative Study: Participants

We recruited 10 professional VFX designers (7 male, 3 female) with professional experience spanning 3 to 26 years (M=10.1, SD=7.0). Participants reported Cinema4D, Unreal Engine, Houdini, Blender, and Unity as their primary tools, with specializations encompassing explosion effects, particle systems, smoke simulations, and magical spell effects. Detailed participant-level profiles are provided in Table 1.

Table 1: Details about the professional VFX designers who participated in the formative study.

ID	Age	Gender	Exp.	Primary Tool	Specialization
P1	26	F	5 yr	Unreal, Cinema4D	Magic, Smoke
P2	31	M	7 yr	Unreal, Houdini	Smoke, Destruction
P3	45	M	10 yr	Cinema4D, Unity	Smoke, Explosions
P4	35	F	10 yr	Cinema4D, Unity	Magic, Explosions
P5	34	M	10 yr	Blender, Rhino	Magic, Simulation
P6	28	M	5 yr	Unreal, Unity	Realistic sim., Flames
P7	43	M	15 yr	Houdini, Blender	Smoke, Explosions
P8	53	M	26 yr	Blender, Unreal	Explosions, Projectiles
P9	24	M	3 yr	Cinema4D, Unreal	Liquid sim., Magic
P10	28	F	3 yr	Houdini, Blender	Motion, Smoke

3.2 Formative Study: Procedure

The interviews explored three main areas: (1) VFX asset design approaches and requirements, (2) spatial-temporal coordination challenges in VFX workflows, and (3) description preferences for expressing VFX concepts. To understand how professionals specify VFX concepts, we presented participants with three reference VFX images and asked them to describe how they would request an AI system to generate similar effects.

3.3 Formative Study: Findings

VFX Meshes and Textures Differ from Conventional 3D Assets. Participants revealed that VFX assets differ from conventional 3D models in both purpose and structure. In traditional 3D modeling, topology and geometry serve visual fidelity: *"Properly structured topology matters because the mesh itself is the final visual output"*(P1). In VFX, however, meshes are only one layer in a multi-component composition, and quality depends more on how sub-components work together: *"I spend most of my time arranging sub-components, not perfecting a single mesh"*(P4). As illustrated in Figure 2, VFX assets are built by composing multiple simple sub-components rather than crafting a single detailed object. Existing 3D generation models, which focus on producing individual high-fidelity meshes and textures, are structurally mismatched with this compositional workflow.

Meshes provide topology rather than geometric detail. VFX meshes serve as topology carriers for shader-driven visual effects, not as detailed geometric representations. Participants explained that meshes need to be simple specifically because visual complexity comes from textures applied on top: *"Simple meshes are used because textures play a large role"*(P7), *"I choose meshes based on general shapes rather than geometric detail"*(P4). This means VFX meshes must support UV manipulation and texture scrolling rather than carry geometric detail, a requirement that conventional 3D generation models do not address.

Textures utilize noise rather than representational imagery. VFX textures also diverge from conventional 3D textures. Rather than representational images mapped to surfaces, VFX relies on tileable, grayscale noise patterns that shaders manipulate to produce dynamic visual effects: *"We use a lot of noise textures, and we mix noise with another noise to create effects"*(P5), *"Procedurally generated noise effects are extensively used"*(P7). These grayscale patterns allow

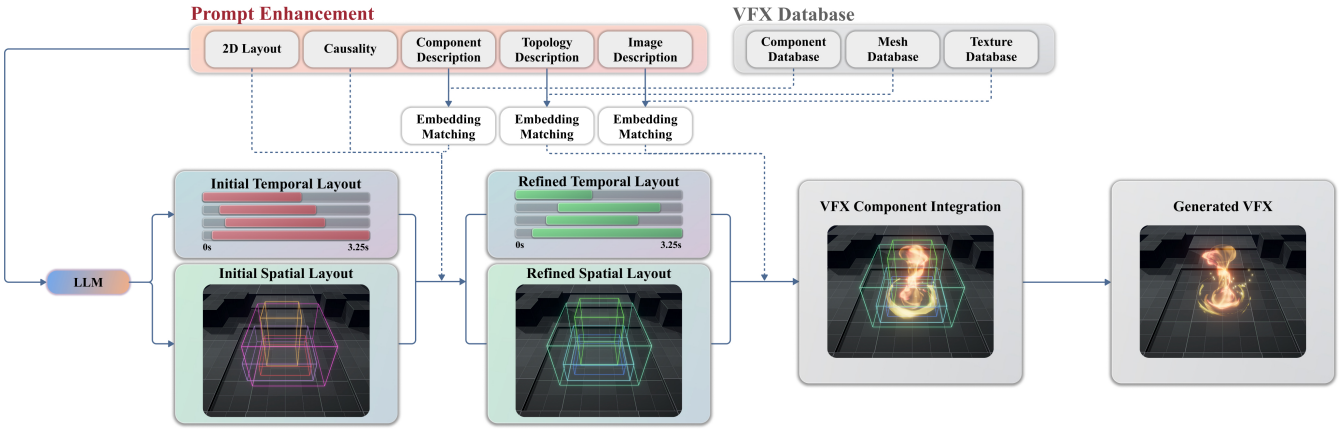


Figure 3: Structure of SpeLL: (1) Prompt enhancement module, (2) Spatial and temporal layout design generation modules, (3) embedding-based retrieval from VFX database, and (4) component integration.

shaders to assign colors and control transparency at runtime, while their tileable structure enables texture scrolling for animation.

Spatial-Temporal Coordination Demands Manual Effort. All participants identified spatial and temporal coordination as the primary aspects of professional VFX design, requiring coordination of multiple sub-components. For example, a ground-cracking magic effect involves sequential sub-effects: ground cracks → debris fragments → smoke plumes → light emissions, each requiring timing and positioning to create cause-and-effect relationships. This coordination requires manual adjustment of spatial positioning and temporal sequencing through iterative trial-and-error processes.

Temporal coordination follows logical sequencing based on physical causality: *"I identify and arrange logical sequential structures between effect elements"*(P1). Spatial arrangement follows a hierarchy that directs visual attention to primary components: *"Most of the strongest elements are placed near the main object"*(P3).

Preferences for Generating Complete Asset with Holistic and Simplified Prompting. When describing VFX for AI generation, participants favored holistic descriptions of overall appearance and function over detailed component-level specifications. Rather than specifying particle counts, mesh topology, or timing parameters, participants described functional context and overall visual appearance. Because the question explicitly framed an AI system, these descriptions capture participants' prompt-based interaction at the time of the study. Participants also expected the resulting structure to remain editable. Participants valued AI for directly generating initial sources: *"I want AI to generate effects that can be directly utilized during work"* (P1), *"It would be good if AI generates VFX layouts. I'd prefer to control myself as a designer"* (P8). Representative prompts for sword slash (*"A trail effect following character sword movement"*(P1)), explosion (*"Large explosion, mushroom-shaped clouds like a nuclear explosion, ground waves spreading in all directions"*(P6)), and healing magic (*"Green dome-shaped effect suitable for healing magic"*(P7)) all demonstrate this preference for simplified, contextual description over technical specification during early exploration.

Particles Serve as Final Enhancement Elements. When asked about elements that significantly enhance VFX quality, most participants identified particle effects as finishing touches: *"Natural particle effects are used for details"*(P1), *"Detailed particle expressions like fragments and dust are used"*(P3). These elements are coordinated to match the movement and texture scrolling of main sub-components, serving as atmospheric details that complete the overall VFX composition.

3.4 Formative Study: Design Objectives

Our formative study reveals that professional VFX creation is fundamentally a sequence of compositional decisions: selecting mesh topologies, pairing them with appropriate noise textures, and arranging sub-components spatially and temporally. Rather than synthesizing new assets through generative models, SpeLL automates an initial pass over these decision making steps by using an LLM to select, retrieve, and arrange existing professional assets. This mirrors how a skilled designer draws on prior references to make compositional decisions rather than creating every element from scratch, while still leaving the resulting structure open to manual refinement. Based on these findings, we define three design objectives:

- **DO1: Support simplified prompting for VFX specification.** SpeLL's prompt enhancement module interprets simplified descriptions and expands them into structured technical specifications including component identification, causality relationships, and spatial arrangements.
- **DO2: Generate production-compatible VFX components.** SpeLL's spatial and temporal design modules retrieve professional-grade meshes, textures, and parameter patterns from a reference database and assemble them into production-compatible VFX assets, mirroring the manual selection and assembly process that designers perform.
- **DO3: Provide complete VFX templates for manual refinement.** SpeLL generates multi-component VFX templates as starting points, providing professionals with structured foundations they can manually adjust and polish.

4 SpeLL

SpeLL transforms natural language descriptions into editable visual effects in Unity Engine through three modules (Figure 3): (1) **Prompt Enhancement Module**, which converts simplified user descriptions into structured specifications; (2) **Spatial Design Module**, which handles component selection, retrieval, and 3D spatial arrangement; and (3) **Temporal Design Module**, which sequences component activation and coordinates timing. SpeLL is intended to support early exploration and draft generation rather than to replace manual authoring or execute a fully specified final effect from text alone.

Rather than synthesizing assets through generative models, SpeLL treats VFX creation as a design decision problem: the LLM proposes which components to include, where to place them, and how to sequence them, and designers can then revise the generated structure through parameter editing. Both design modules retrieve pre-existing assets from a VFX database of 1,502 professionally authored assets from Unity's Asset Store (decomposed into 18,010 sub-components, 261 unique meshes, and 217 unique textures), functioning as a reference library from which a designer would manually select and combine components. To comply with Unity Asset Store's licensing terms, no model is trained or fine-tuned on these assets; all LLMs and embedding models operate in inference mode only, and retrieved assets are procedurally combined and transformed at assembly time according to the LLM's layout decisions. For indexing, VFX components are described by rendering temporal frames and passing them to Gemini-2.5-Flash for text description. Meshes are described via multi-view renders and LLM-generated topology descriptions, and textures are described using VFX and component names as context. All descriptions are encoded into 384-dimensional embedding vectors via a pre-trained Sentence Transformer [46] for similarity-based retrieval.

4.1 Prompt Enhancement Module

The prompt enhancement module converts simple user prompts into structured inputs suitable for VFX generation. Following formative-study findings, SpeLL preserves natural prompting while extracting actionable design constraints for downstream modules. Concretely, the module generates a visual reference and then derives component semantics, spatial hints, and temporal cues from that reference. The synthesized image is used as an intermediate scaffold for reasoning about components and layout.

Image Synthesis for Design Reference. The first stage uses gpt-image-1 to synthesize a visual reference from the user prompt (Figure 4). This reference externalizes sub-components that are difficult to specify with text only and provides a shared artifact for subsequent spatial-temporal parsing.

Image Analysis with LLM. Gemini-2.5-Flash then extracts three outputs: (1) component-level semantic descriptions for retrieval, (2) 2D layout cues (bounding boxes) for initializing 3D placement, and (3) temporal dependency cues for sequencing component activation. Together, these outputs become structured constraints for the spatial and temporal design modules.

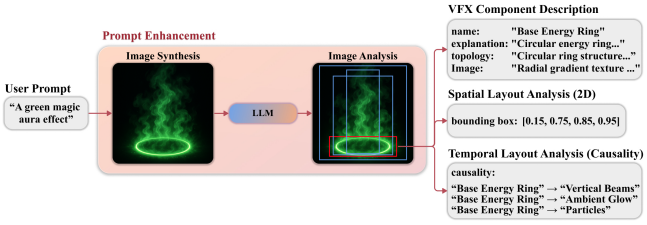


Figure 4: Prompt Enhancement module: (1) image synthesis generates a visual reference, (2) LLM analysis extracts component descriptions, spatial layouts, and temporal causality.

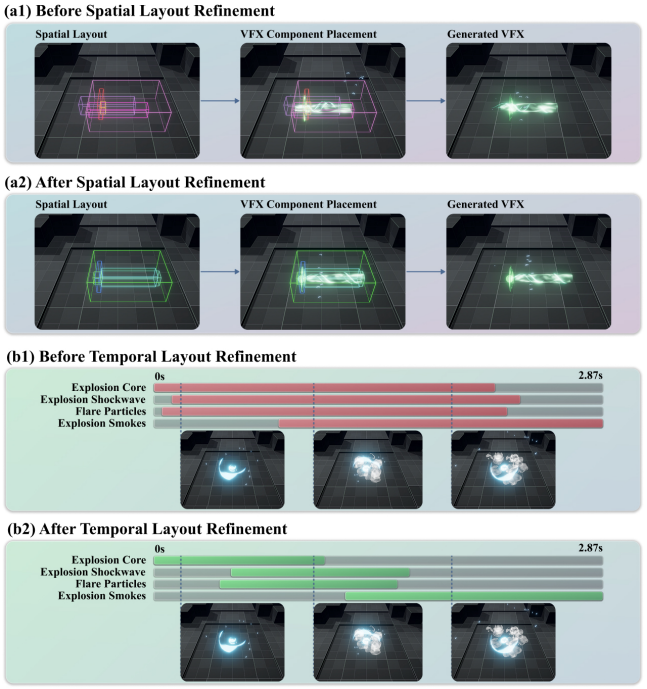


Figure 5: (a1-a2) Spatial refinement transforms misaligned component positions into properly arranged layouts. (b1-b2) Temporal refinement reorganizes component timing with unclear causality to logical relationships.

4.2 Spatial Design Module

The spatial design module determines VFX component selection and 3D spatial arrangement (Figure 5). Given structured specifications from the prompt enhancement module, it retrieves similar reference components and assets from the VFX database, uses an LLM to generate an initial layout, and refines component positions through optimization.

VFX Component, Mesh, and Texture Retrieval. SpeLL performs three embedding-based searches using the component descriptions from image analysis. VFX component retrieval supplies spatial and temporal parameter references, while mesh and texture retrieval supplies candidate assets for constructing each component.

Initial Spatial Layout Generation. A unified Gemini-2.5-Flash design request generates the initial spatial-temporal VFX specification, including 3D bounding boxes for component positions and scales. These spatial parameters, combined with 2D layout references from the prompt enhancement stage, provide the starting configuration. However, direct application of LLM-generated parameters can produce spatially misaligned results, and the 2D reference alone cannot resolve depth, necessitating the refinement that follows.

Spatial Layout Refinement. For each component, the stored spatial parameter ratios from the top-N retrieved VFX components serve as reference values for refinement. Figure 5 (a1-a2) illustrates how spatial refinement corrects misaligned component positions. The optimized variables are the layout parameters of each component (positions and sizes), not any model weights. The cost function balances three soft constraints:

$$\mathcal{L}_{\text{spatial}} = \lambda_{\text{init}}\mathcal{L}_{\text{init}} + \lambda_{\text{proj}}\mathcal{L}_{\text{proj}} + \lambda_{\text{database}}\mathcal{L}_{\text{database}} \tag{1}$$

We normalized the loss terms to comparable numerical ranges so that no term dominated because of its raw scale. $\mathcal{L}_{\text{init}} = \sum_i \|\mathbf{p}_i - \mathbf{p}_i^{\text{init}}\|^2 + \|\mathbf{s}_i - \mathbf{s}_i^{\text{init}}\|^2$ anchors optimization to the LLM's initial layout, where $\mathbf{p}_i$ and $\mathbf{s}_i$ are the position and size of component i. This term preserves coarse spatial ordering from the LLM proposal when projection constraints are underdetermined. $\mathcal{L}_{\text{proj}}$ enforces consistency between the 3D layout and the 2D bounding boxes from prompt enhancement. For each component, all 8 corners of the 3D bounding box are projected to 2D via the view matrix $\mathbf{V}(\theta)$ and perspective projection matrix $\mathbf{P}$:

$$\mathcal{L}_{\text{proj}} = \sum_i \|\mathbf{c}_i^{\text{proj}}(\theta) - \mathbf{c}_i^{\text{target}}\|^2 + \|\mathbf{s}_i^{\text{proj}}(\theta) - \mathbf{s}_i^{\text{target}}\|^2 \tag{2}$$

where $\mathbf{c}_i^{\text{proj}}$ and $\mathbf{s}_i^{\text{proj}}$ are the projected 2D center and size, jointly optimizing component positions and camera pose θ. Since this single-view projection does not fully constrain depth, $\mathcal{L}_{\text{database}}$ provides an additional regularizer by pulling each component's normalized size toward its embedding-weighted reference average:

$$\mathcal{L}_{\text{database}} = \sum_i \left(\textit{norm-size}_i - \overline{\textit{norm-size}}_{\text{ref}(i)}\right)^2 \tag{3}$$

Together, the three terms use the LLM proposal for coarse ordering, the reference image for visible composition, and retrieved examples for component proportions. The result remains one plausible 3D layout rather than a unique reconstruction, and designers can further adjust component transforms in the edit panel.

4.3 Temporal Design Module

The temporal design module sequences component activation and coordinates timing relationships. It generates initial temporal layouts via the LLM, uses professional temporal patterns from the retrieved VFX components, and refines timing through optimization.

Initial Temporal Layout Generation. The same initial spatial-temporal VFX design request generates delay and duration parameters for each component, specifying when each sub-component appears and how long it persists. However, these initial temporal arrangements exhibit deficiencies due to the LLM's limited understanding of VFX timing design, requiring the refinement that follows.

Temporal Layout Refinement. Temporal refinement adjusts the LLM's initial timing using causality relationships from the prompt enhancement stage and professional temporal patterns retrieved from the VFX database. Figure 5 (b1-b2) illustrates how temporal refinement corrects misaligned component timings. The optimized variables are the delay and duration parameters of each component, not any model weights. The cost function balances four terms:

$$\begin{aligned}\mathcal{L}_{\text{temporal}} = \lambda_{\text{init}}\mathcal{L}_{\text{init}} + \lambda_{\text{causality}}\mathcal{L}_{\text{causality}} \\ + \lambda_{\text{coverage}}\mathcal{L}_{\text{coverage}} + \lambda_{\text{database}}\mathcal{L}_{\text{database}}\end{aligned} \tag{4}$$

We normalized the four loss terms to comparable numerical ranges so that no term dominated because of its raw scale. $\mathcal{L}_{\text{init}} = \sum_i (d_i - d_i^{\text{init}})^2 + (\tau_i - \tau_i^{\text{init}})^2$ preserves the LLM's original timing, where d_i and τ_i are the delay and duration of component i. $\mathcal{L}_{\text{causality}}$ enforces the causal and concurrent ordering relationships extracted by the prompt enhancement module:

$$\mathcal{L}_{\text{causality}} = \sum_{(i,j)\in\mathcal{R}_{\text{causal}}} \mathcal{L}_{\text{causal}}(i,j) + \sum_{(i,j)\in\mathcal{R}_{\text{concurrent}}} \mathcal{L}_{\text{concurrent}}(i,j) \tag{5}$$

where $\mathcal{L}_{\text{causal}}(i,j) = (t_i^{\text{end}} - t_j^{\text{start}})^2$ when $t_i^{\text{end}} > t_j^{\text{start}}$ (zero otherwise), and $\mathcal{L}_{\text{concurrent}}(i,j) = (t_i^{\text{start}} - t_j^{\text{start}})^2 + (t_i^{\text{end}} - t_j^{\text{end}})^2$. $\mathcal{L}_{\text{coverage}}$ penalizes time gaps where no component is active by sampling N uniform points over $[0, T_{\text{total}}]$. $\mathcal{L}_{\text{database}}$ aligns each component's normalized temporal midpoint toward its embedding-weighted reference average:

$$\mathcal{L}_{\text{database}} = \sum_i \left(\textit{norm-midpoint}_i - \overline{\textit{norm-midpoint}}_{\text{ref}(i)}\right)^2 \tag{6}$$

4.4 VFX Component Assembly

Following layout refinement, the system assembles the final VFX from the retrieved meshes and textures and configures particle dynamics from the optimized layouts. To avoid forcing semantically weak matches, SpeLL only instantiates retrieved components when their embedding similarity exceeds a predefined threshold; otherwise, the unreliable component is omitted from the generated draft. Designers can then add missing components manually through the component hierarchy and assign meshes, textures, and parameters using the editing interface.

Procedural Mesh Synthesis. Mesh synthesis combines the N retrieved database meshes in three stages. Vertices are first globally scaled to match embedding-weighted average reference proportions, then locally deformed via closest-surface projection:

$$\mathbf{v}_i^{\text{deformed}} = \mathbf{v}_i^{\text{scaled}} + \sum_{j=1}^{N} w_j \cdot s_j \cdot (\mathbf{c}_j(\mathbf{v}_i^{\text{scaled}}) - \mathbf{v}_i^{\text{scaled}}) \tag{7}$$

where $\mathbf{c}_j(\mathbf{v})$ is the closest surface point on reference mesh j and s_j is a strength factor. Finally, iterative Laplacian smoothing removes deformation artifacts while preserving continuity across seams in the mesh's UV texture coordinates.

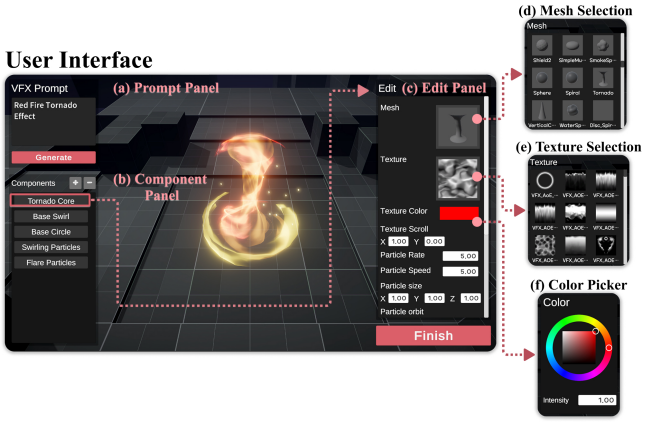


Figure 6: SpeLL user interface: (a) Prompt Panel, (b) Component Panel, (c) Edit Panel, (d) Mesh Selection, (e) Texture Selection, and (f) HDR Color Picker.

Table 2: Parameters available in the VFX creation interface.

Category	Adjustable Parameters
Temporal Layout	Delay; Duration
Spatial Layout	Position (X, Y, Z); Rotation (X, Y, Z); Scale (X, Y, Z)
Emission	Particle Rate; Emission Shape; Emission Radius
Particle Dynamics	Velocity; Orbital (X, Y, Z); Rotation over Lifetime (X, Y, Z); Particle Size (X, Y, Z)
Appearance	Texture Selection; Texture Scroll (X, Y); Mesh Selection; Color Tint (RGB)

Procedural Texture Synthesis. Textures are blended through weighted multiplicative compositing, using embedding similarity as weights for both color and alpha channels across the N retrieved references.

Particle Effects Configuration. Particle velocity is set to traverse the bounding box along the LLM-specified emission direction $\hat{\mathbf{d}}_{\text{emit}}$. When the main sub-component's texture scroll velocity $\mathbf{t}_{\text{scroll}}$ produces tangential motion ($\mathbf{v}_{\text{tangent}} = \mathbf{t}_{\text{scroll}} - (\mathbf{t}_{\text{scroll}} \cdot \hat{\mathbf{d}}_{\text{emit}})\hat{\mathbf{d}}_{\text{emit}}$) and the mesh forms a closed topology in the scroll direction, orbital velocity $|\mathbf{v}_{\text{tangent}}| \cdot \hat{\mathbf{n}}_{\text{orbit}}$ is added to synchronize particle motion with the main component's texture animation.

4.5 User Interface

SpeLL provides a web-based interface built with Unity WebGL (Figure 6). Users enter natural language descriptions in the prompt panel (Figure 6(a)) to generate VFX, then browse the component hierarchy (Figure 6(b)) and adjust individual sub-component parameters in the edit panel (Figure 6(c)), including mesh assignment (Figure 6(d)), texture selection (Figure 6(e)), and HDR color (Figure 6(f)). These controls allow artists to treat generated results as editable drafts: they can adjust transforms, delay and duration, particle dynamics, emission properties, color, texture scrolling, and mesh or texture selection after generation. Table 2 lists the complete set of adjustable parameters.

Table 3: Runtime breakdown across 30 prompts collected during the formative study. Values report mean seconds with sample standard deviations.

Process	Time (s), *M* (*SD*)
Reference image synthesis	25.93 (8.75)
LLM image analysis	4.66 (1.28)
VFX component, mesh, and texture retrieval	1.23 (0.74)
Initial spatial-temporal VFX design	2.71 (0.75)
Spatial-temporal layout optimization	0.65 (0.17)
Unity VFX hierarchy instantiation	0.22 (0.04)
Total generation time	35.47 (8.86)

4.6 Runtime Performance

We measured SpeLL's VFX generation time using the 30 prompts collected during the formative study, comprising one description from each of the 10 participants for each of the three reference VFX. Measurements were collected after application initialization on an Apple M1 processor. Table 3 reports the runtime of each process.

5 MAIN STUDY

We conducted a user study with three tasks to evaluate whether SpeLL achieves its design objectives by comparing it with alternative approaches.

5.1 Main Study: Participants

We recruited 20 professional VFX designers (8 male, 12 female, ages 24-38, M=26.1, SD=3.7) with at least three years of experience in game development and visual effects production. Participants had extensive experience with Unity engine (M=3.4 years, SD=2.1) and reported creating VFX for multiple game genres including action, RPG, and mobile games. None of the main-study participants participated in the formative study.

5.2 Main Study: Procedure

The study consisted of three tasks to validate our design objectives: Task 1 (Ablation Study) validated DO1 by comparing SpeLL against a framework without prompt enhancement; Task 2 (Generation Module Comparison Study) validated DO2 by comparing SpeLL's LLM-based generation against direct 3D synthesis; Task 3 (Manual Workflow Comparison Study) validated DO3 by comparing AI-assisted and manual VFX creation.

We deployed the study interface as a WebGL-built application that participants accessed remotely through web browsers. Participants were not informed about which framework they were using during any of the tasks; conditions were presented through the same interface without framework labels. For validation, we selected three VFX scenarios based on findings from our formative study. Professional VFX designers reported typically receiving requests that specify the purpose and general appearance of effects rather than detailed technical specifications. They also identified magic circles, explosions/smoke, and projectiles as the most frequently designed VFX categories in their practice. Accordingly, we assigned participants to create: a green healing magic effect for character restoration (Task 1), an explosion with smoke effect for

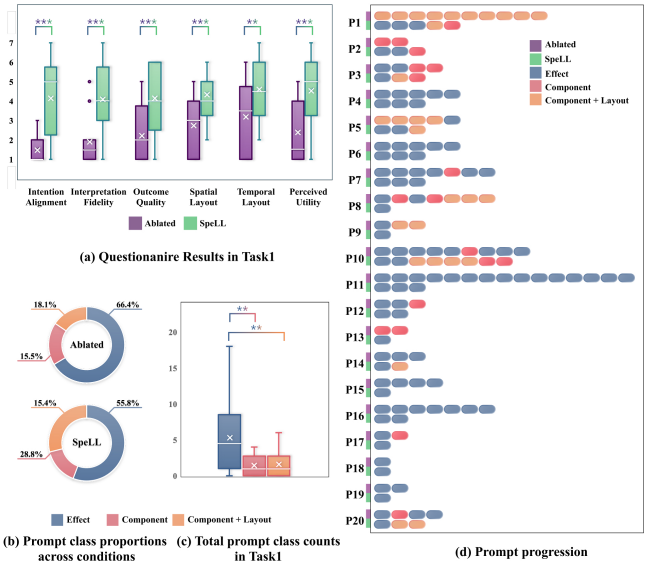


Figure 7: Task 1 quantitative results. (a) Questionnaire scores across six evaluation dimensions. (b)&(c) Distribution of prompt types by description granularity. (d) Prompt type progression across prompt order for each participant.

destruction scenes (Task 2), and a fire projectile effect for flame magic (Task 3).

5.3 Task 1: Ablation Study (DO1)

To validate DO1, which aims to support simplified prompting for VFX specification, we investigated the contribution of SpeLL's prompt enhancement module by designing an ablated framework that bypasses visual reference generation and component analysis. This comparison evaluates how prompt enhancement affects both the quality of generated VFX assets and the patterns of user interaction with the system by comparing *SpeLL* and *Ablated* frameworks.

Task 1: Ablated Framework. The *Ablated* framework removes the prompt enhancement module, requiring the spatial and temporal design modules to handle all aspects of VFX generation based only on the raw text description, placing the entire burden of interpreting user intent on the language model's inherent capabilities. Since 2D bounding boxes and causality relationships from prompt enhancement are unavailable, $\mathcal{L}_{\text{proj}}$ is excluded from the spatial loss and $\mathcal{L}_{\text{causality}}$ is excluded from the temporal loss compared to the full SpeLL formulation. Full loss formulations are provided in Appendix A.

Task 1: Measures and Questionnaires. To evaluate user experience and performance between the *Ablated* and *SpeLL* frameworks, we developed a questionnaire measuring six key dimensions using 7-point Likert scales, as shown in Appendix–Table 4. Additionally, we collected all prompting interactions to analyze how participants engaged with each framework.

To analyze prompting patterns, we classified collected prompts into three categories by the granularity of detail provided. All three

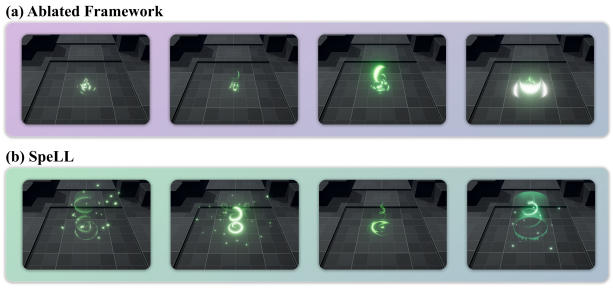


Figure 8: Representative VFX outputs from Task 1 created by participants. (a) Ablated Framework outputs, (b) SpeLL outputs.

levels describe the desired VFX as a whole, but differ in how much internal structure the user specifies: *Effect-level* prompts describe only the overall appearance (e.g., *"a green healing aura"*), *Component-level* prompts additionally specify individual sub-components (e.g., adding *"with floating leaf particles and a base ring"*), and *Layout-level* prompts further detail spatial relationships among those sub-components (e.g., adding *"ring at the bottom, light columns ascending above"*). We conducted repeated measures ANOVA to assess differences among these prompt types, followed by post-hoc tests with Bonferroni correction.

Task 1: Results. We assessed the normality of all measures using the Shapiro-Wilk test. For measures that satisfied normality assumptions, we conducted paired samples t-tests to assess significance between the *Ablated* and *SpeLL* conditions. For measures that violated normality assumptions, we used Wilcoxon signed-rank tests. Significance levels are indicated as $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$. Detailed statistics are presented in Appendix–Table 7 (a) and (b).

Improvements across all evaluation measurements. As illustrated in Figure 7(a), *SpeLL* significantly outperformed the *Ablated Framework* across all six dimensions. The largest gains were in intention alignment, with *SpeLL* ($M = 4.15, SD = 1.76$) significantly outperforming *Ablated* ($M = 1.60, SD = 0.82$, $p < .001^{***}$), and perceived utility, with *SpeLL* ($M = 4.45, SD = 1.85$) over *Ablated* ($M = 2.40, SD = 1.64$, $p < .001^{***}$); all remaining dimensions also reached significance (all $p \leq .006$). Representative outputs are shown in Figure 8.

Participants begin with effect-level prompts. As shown in Figure 7(b), professional designers predominantly used effect-level descriptions. For example, *"soft green healing aura with divine light"*(P3) was classified as *effect-level*, *"green particles with floating leaf shapes, base green ring"*(P7) as *component-level*, and *"circular healing effect at the bottom with ascending light columns on top of the rings, and cross sign at the center"*(P12) as *layout-level*. Repeated measures ANOVA showed significant differences in usage frequency ($F(1.23, 23.39) = 7.638, p = .008^{**}$). Post-hoc tests demonstrated that participants used *effect-level* prompts significantly more frequently than *component-level* prompts ($p = .004^{**}$) or *layout-level* prompts ($p = .006^{**}$), with no significant difference between the latter two ($p = 1.000$). The result indicates that, in this

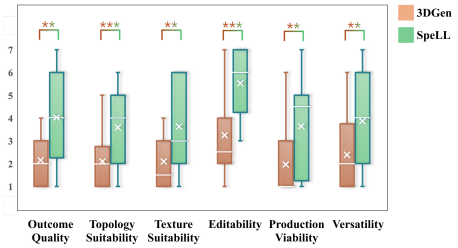


Figure 9: Questionnaire scores across six evaluation dimensions measured in Task 2.

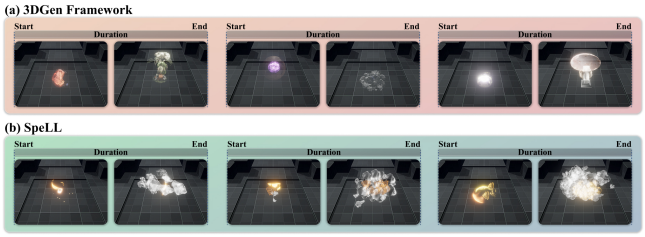


Figure 10: Representative VFX outputs from Task 2 created by participants for the explosion and smoke effect. (a) 3DGen outputs, (b) SpeLL outputs.

early-exploration task setting, professional VFX designers often began with effect-level descriptions and added more specific details in subsequent prompts. The prompt type progression across participants is shown in Figure 7(d).

5.4 Task 2: Generation Method Comparison Study (DO2)

Our formative study revealed that VFX assets require simple meshes and noise textures rather than the geometrically complex outputs of 3D generation models. To validate DO2 and empirically test whether this structural mismatch affects professional workflows, we replaced SpeLL's spatial and temporal design modules with Trellis, a state-of-the-art text-to-3D model [60], comparing retrieval-based assembly against direct 3D synthesis.

Task 2: 3D Generation Framework. The 3D Generation Framework (*3DGen*) maintains SpeLL's prompt enhancement module but directly synthesizes 3D meshes from component descriptions via Trellis, then optimizes spatial and temporal layouts based on prompt enhancement outputs. The framework derives initial spatial positions by mapping 2D bounding box centers from image analysis to 3D world coordinates, and derives initial temporal delays from the causality order of each component. The loss structure retains the same terms as SpeLL's full formulation, but $\mathcal{L}_{\text{init}}$ is anchored to these prompt-derived layouts rather than LLM outputs. Full formulations are provided in Appendix A.

Task 2: Questionnaires. To evaluate the performance differences between the *3DGen* and *SpeLL* frameworks, we used a questionnaire measuring six dimensions using 7-point Likert scales, as shown in Appendix–Table 5. The questionnaire focused on assessing how well each generation approach serves professional VFX production requirements given their different output formats.

Task 2: Results. We assessed normality using the Shapiro-Wilk test, then applied paired samples t-tests or Wilcoxon signed-rank tests as appropriate. Significance levels are indicated as $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$. Complete statistics are provided in Appendix–Table 7 (c). As illustrated in Figure 9, *SpeLL* outperformed *3DGen* across all six dimensions. Editability showed the largest gap, with *SpeLL* ($M = 5.40, SD = 1.40$) outperforming *3DGen* ($M = 3.25, SD = 2.00$, $p < .001^{***}$); all other dimensions similarly favored *SpeLL* (all $p \leq .002$). These results indicate that *SpeLL* produces VFX in formats that align with professional production workflows. Representative outputs are shown in Figure 10.

5.5 Task 3: Manual Workflow Comparison Study (DO3)

To validate DO3, which aims to provide complete VFX templates for manual refinement, we investigated how AI-generated templates affect the creative workflow compared to traditional approaches. We designed a *Manual* creation interface that provides conventional VFX authoring capabilities without AI support. This comparison evaluates user experience and creative satisfaction between traditional manual workflows and AI-assisted approaches that offer initial templates as starting points.

To ensure a controlled comparison, both conditions shared identical design spaces: the *Manual* condition provided access to the same mesh and texture databases, while SpeLL's procedural mesh and texture synthesis modules were disabled so all outputs could be reproduced manually. Participants in the *Manual* condition assembled components and configured parameters directly; in the *SpeLL* condition, they generated an AI-assisted draft via text prompt and then refined it through the same parameter manipulation controls. Participants were not required to define a fixed artistic target in advance; their intent could be formed or revised as they inspected intermediate results.

Task 3: Questionnaires. To evaluate the creative workflow benefits of AI-assisted VFX design, we employed the Creativity Support Index (CSI) [15] for measuring creativity support in digital tools. We adapted CSI to our VFX domain by selecting five measurements relevant to creative VFX workflows: enjoyment, exploration, expressiveness, immersion, and results worth effort. Each dimension was measured using two questionnaire items on 10-point Likert scales, as shown in Appendix–Table 6.

Task 3: Results. We assessed normality using the Shapiro-Wilk test, then applied paired samples t-tests or Wilcoxon signed-rank tests as appropriate. Significance levels are indicated as $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$. Complete statistics are provided in Appendix–Table 7 (d).

Significant differences across creativity support dimensions. As illustrated in Figure 11, *SpeLL* outperformed *Manual* across four of five creativity dimensions. Enjoyment showed *SpeLL* ($M = 6.20, SD = 2.23$) outperforming *Manual* ($M = 3.75, SD = 2.09$, $p = .001^{**}$). Exploration favored *SpeLL* ($M = 6.25, SD = 2.26$) over *Manual* ($M = 4.45, SD = 2.68$, $p = .007^{**}$). Results worth effort showed *SpeLL* ($M = 5.83, SD = 2.66$) compared to *Manual*

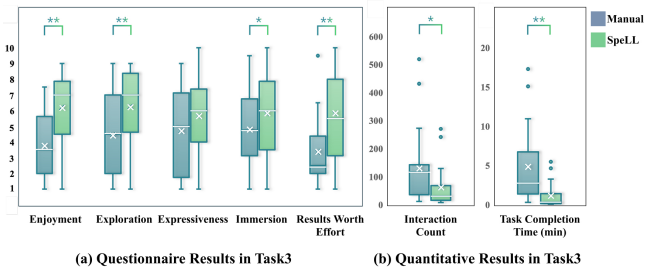


Figure 11: Task 3 quantitative results. (a) Questionnaire scores across five creativity support dimensions, (b) Interaction count, (c) Task completion time.

($M = 3.38, SD = 2.15, p = .001^{**}$). Immersion demonstrated *SpeLL* ($M = 5.83, SD = 2.52$) over *Manual* ($M = 4.80, SD = 2.36, p = .048^{*}$). Expressiveness showed no significant difference ($p = .084$), indicating that SpeLL reduced effort and supported exploration but did not exceed the expressive range of manual authoring. Representative outputs are shown in Figure 12.

Reduced task completion time and interaction count. *SpeLL* significantly reduced both task completion time and interaction count compared to manual creation. Task completion time was shorter with *SpeLL* ($M = 1.16, SD = 1.59$ min) than *Manual* ($M = 4.77, SD = 4.93$ min, $p = .02^{*}$). Interaction count was also lower with *SpeLL* ($M = 61.70, SD = 74.19$) compared to *Manual* ($M = 129.50, SD = 136.23, p = .002^{**}$). These quantitative measures indicate that AI-assisted template generation substantially reduces the effort required to produce VFX assets. Detailed statistics are provided in Appendix–Table 7.

6 DISCUSSION

6.1 Simplified Prompting and Iterative Feedback Loops (DO1)

In the early-exploration task, participants often began with concise, effect-level prompts and moved to specific refinements after inspecting a generated draft. In interviews, they noted that SpeLL understood their high-level intent: *"The AI understood basic components and colors from my prompts"* (P4). This allowed them to focus on conceptual description rather than implementation: *"I could focus on overall visual impressions rather than specifying particle counts or timing parameters"* (P1). SpeLL thus supports a progression from high-level ideas to manual refinement once an editable structure exists.

The prompt progression data (Figure 7(d)) further reveals how this preference shaped the iterative feedback loop within each session. Under SpeLL, participants typically began with an *effect-level* prompt. When they continued, subsequent prompts were reactive rather than speculative: they responded to concrete issues in the generated output. P10, for instance, issued eight prompts, each targeting a specific problem in the preceding result: *"remove the surrounding circle effect"*, then *"make the particles more like small leaves"*, then *"disperse them more irregularly"*. Each prompt addressed what the system had just produced, rather than describing a still-imagined effect from scratch.

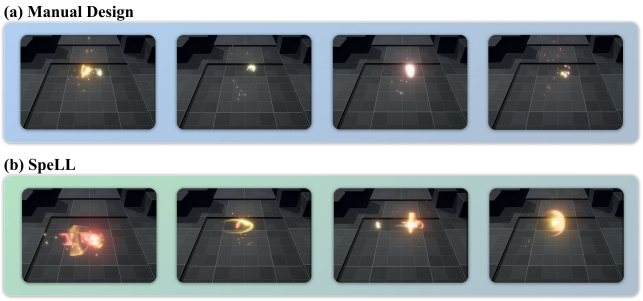


Figure 12: Representative VFX outputs from Task 3 created by participants for the fire projectile effect. (a) Manual outputs, and (b) SpeLL outputs.

The Ablated condition produced a different prompting loop. Some participants resubmitted near-identical *effect-level* prompts with minor wording changes: P11 issued 15 prompts, many semantically similar. Others began directly with *layout-level* prompts: P1 submitted all ten prompts at this level. Prompting in the Ablated condition therefore alternated between repeated high-level reformulation and early specification of component relationships, whereas SpeLL more often supported refinement in response to a generated draft.

This contrast between SpeLL and the Ablated condition aligns with a broader trend in LLM-based spatial understanding. Earlier methods typically require users to specify detailed relationships between objects [20, 40, 64], while recent approaches have begun exploring simpler prompting that lowers user burden and encourages design exploration [21, 23, 33, 35, 62]. Our findings provide empirical evidence from professional practice that VFX workflows similarly benefit from systems that interpret high-level conceptual input, enabling the reactive refinement loop observed in the SpeLL condition.

6.2 Production-Compatible VFX Component Generation (DO2)

Task 2 results support our formative study observation that VFX assets have structurally different requirements from general 3D objects, and that direct 3D generation does not readily satisfy these requirements. Participants identified mismatches at both the mesh and texture level. For meshes, 3D generation models tended to produce geometrically complex outputs that lacked the simple topology VFX workflows rely on: *"VFX meshes need to support panning, rotation, and distortion through UV manipulation"* (P8), *"3D generation creates static props, not effective for VFX design"* (P11). For textures, the representational imagery generated by 3D models was less compatible with the noise-based, shader-driven approach used in VFX production: *"The generated textures looked realistic but couldn't be scrolled or tiled the way we need"* (P4). At the complete asset level, the lack of decomposition into separately editable subcomponents made 3D-generated VFX less practical for iterative refinement: *"Without parameter control, I'd need to recreate everything from scratch anyway"* (P11). These observations are consistent with professional AI adoption research showing that practitioners tend to prioritize tools whose outputs integrate into existing

pipelines [6, 14, 41, 42, 57], and suggest that VFX generation may benefit from producing parameterized, decomposed assets rather than monolithic 3D outputs [3].

6.3 AI-Generated Templates as Starting Points for Refinement (DO3)

Task 3 evaluated whether SpeLL's template generation (DO3) provides suitable starting points for professional refinement. Participants noted that AI-generated outputs allowed them to focus on creative decisions rather than technical setup: *"Having an initial design from AI made me think bigger"* (P20), *"When AI handles the tedious parts, I set higher quality targets for the final output"* (P8). This aligns with research showing that generative AI provides functional baselines that reduce exploration costs [7, 11, 52]. Automated layout coordination freed designers to focus on refinement: *"The AI handled timing between explosion and smoke naturally"* (P5), *"AI can guarantee a certain level of quality so I can concentrate on finalizing details"* (P6). This preference for modifiable initial results indicates that AI tools benefit from preserving professional control when integrating into existing workflows [2, 34, 63].

While SpeLL improved most measurements in Task 3, expressiveness showed no significant difference between conditions ($p = .084$). SpeLL therefore improved the efficiency of reaching an editable structure without expanding the expressive range available through manual authoring. Similar trade-offs have been reported in prior studies of generative AI, where gains in productivity or output quality were accompanied by lower average novelty or more convergent outcomes [7, 69]. Participants' responses revealed that professionals already possess well-developed personal workflows to express their designs: *"I could implement the exact VFX from my imagination"* (P6 on manual condition). Some professionals reported feeling constrained when working with AI-generated foundations: *"Since results are modified from generated outputs, I can express within defined boundaries"* (P5).

Across the three tasks, SpeLL retained familiar VFX controls over component hierarchies, transforms, timing, particle dynamics, appearance, and asset selection, while shifting effort from initial assembly to refinement of an AI-generated structure. Generated drafts helped make partial design intentions concrete, but the synthesized reference image or retrieved components could also anchor subsequent choices. SpeLL therefore keeps each result decomposed and editable, allowing designers to add, remove, replace, retime, or stylize components while retaining control over final expressive decisions [54, 69].

7 LIMITATIONS

SpeLL provides less guidance for hyper-stylized effects and studio-specific visual languages that are underrepresented in its Unity Asset Store database: *"Custom asset import would make this production-ready"* (P10). Its physics-based causality constraints can further restrict stylized effects when artistic intent diverges from realistic behavior: *"Physical causality is good for realism but limits stylized effects"* (P15). User-imported assets, few-shot retrieval augmentation, and adjustable or style-aware causality constraints could help SpeLL accommodate a broader range of studio-specific and stylized VFX.

SpeLL's performance also depends on the image synthesis model's capabilities: when synthesis outputs differ from user intent, downstream modules inherit these misalignments. Participants noted: *"The system couldn't capture the specific magical effect style I was trying to create"* (P7). Incorporating user-in-the-loop editing of the synthesized reference image, or leveraging controllable generation models that accept spatial and stylistic constraints, could improve alignment between user intent and the visual reference that drives subsequent design stages.

Sketch and notation based systems demonstrate how complementary input can support iterative image and scene editing [30, 49, 51]. SpeLL could similarly let designers supply or edit reference images, compare multiple candidates, and revise enhanced specifications or spatial-temporal layouts before retrieval and assembly. These checkpoints could reduce cascading errors while giving designers clearer control over how their input is interpreted.

8 CONCLUSION

This study introduced SpeLL, a framework for generating editable 3D VFX drafts from text prompts using large language models. Through formative studies with professional VFX designers, we identified three design requirements: preference for simplified prompting, need for VFX assets, and manual spatial-temporal coordination challenges. SpeLL addresses these requirements through prompt enhancement, a spatial design module for component retrieval and 3D arrangement, and a temporal design module for activation sequencing and timing coordination. Our user study with twenty professional designers demonstrated that SpeLL improved intention alignment, spatial-temporal layout quality, editability, and workflow efficiency compared with ablated frameworks, 3D generation baselines, and manual workflows, while expressiveness remained comparable to manual authoring. The results show that SpeLL changes where professional effort is spent: it accelerates the move from high-level intent to an editable structure, while designers retain control over final expressive decisions through component-level refinement. Through this evaluation, we observed that combining general LLM knowledge with domain-specific professional databases enables practical VFX draft generation for early exploration without abandoning the parameterized, editable workflows that professionals rely on.

Acknowledgments

This work was supported by National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (RS-2025-16902996, 50%; RS-2024-00343397, 25%) and by the "Project for Science and Technology Opens the Future of the Region" program, funded by the Ministry of Science and ICT (MSIT) and Gwangju Metropolitan City, Republic of Korea, in 2026 (Project: Convergence Culture Virtual Studio for Realizing an Artificial Intelligence-Based Metaverse, 25%).

References

[1] A'Aeshah Alhakamy and Mihran Tuceryan. 2020. Real-time illumination and visual coherence for photorealistic augmented/mixed reality. *ACM Computing Surveys (CSUR)* 53, 3 (2020), 1–34.

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019.

Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.
[3] Nantheera Anantrasirichai and David Bull. 2022. Artificial intelligence in the creative industries: a review. *Artificial intelligence review* 55, 1 (2022), 589–656.
[4] Rahul Arora, Rubaiat Habib Kazi, Danny M Kaufman, Wilmot Li, and Karan Singh. 2019. Magicalhands: Mid-air hand gestures for animating in vr. In *Proceedings of the 32nd annual ACM symposium on user interface software and technology*. 463–477.
[5] Piyush Bagad, Makarand Tapaswi, and Cees GM Snoek. 2023. Test of time: Instilling video-language models with a sense of time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2503–2516.
[6] Andrew Begemann and James Hutson. 2024. Empirical insights into AI-assisted game development: A case study on the integration of generative AI tools in creative pipelines. *Metaverse* 5, 2 (2024).
[7] Léonard Boussioux, Jacqueline N Lane, Miaomiao Zhang, Vladimir Jacimovic, and Karim R Lakhani. 2024. The crowdless future? Generative AI and creative problem-solving. *Organization Science* 35, 5 (2024), 1589–1607.
[8] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. 2024. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*.
[9] Shivam Chandhok. 2024. Scenegpt: A language model for 3d scene understanding. *arXiv preprint arXiv:2408.06926* (2024).
[10] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14455–14465.
[11] Liuqing Chen, Yaxuan Song, Jia Guo, Lingyun Sun, Peter Childs, and Yuan Yin. 2025. How generative AI supports human in conceptual design. *Design Science* 11 (2025), e9.
[12] Yanlin Chen and Xiwen Chen. 2024. Exploration of Deep Semantic Analysis and Application of Video Images in Visual Communication Design Based on Multimodal Feature Fusion Algorithm. *International Journal of Advanced Computer Science & Applications* 15, 8 (2024).
[13] Yun-Chun Chen, Selena Ling, Zhiqin Chen, Vladimir G Kim, Matheus Gadelha, and Alec Jacobson. 2024. Text-guided controllable mesh refinement for interactive 3d modeling. In *SIGGRAPH Asia 2024 Conference Papers*. 1–11.
[14] Vasilka Chergarova, Mel Tomeo, Heidi Morgan, and Sandra Andrade. 2024. Integrating artificial intelligence (AI) into game development to elevate diverse gameplay elements. *Issues in Information Systems* 25, 2 (2024).
[15] Erin Cherry and Celine Latulipe. 2014. Quantifying the creativity support of digital tools through the creativity support index. *ACM Transactions on Computer-Human Interaction (TOCHI)* 21, 4 (2014), 1–25.
[16] Jorge Chueca, Carlos Cetina, Oscar Pastor, and Jaime Font. 2024. Search-based co-creation of software models: The case of particle systems for video games. *Information and Software Technology* 171 (2024), 107466.
[17] Antonio Anastasio Bruto da Costa and Pallab Dasgupta. 2021. Learning temporal causal sequence relationships from real-time time-series. *Journal of Artificial Intelligence Research* 70 (2021), 205–243.
[18] Faraz Faruqi, Ahmed Katary, Tarik Hasic, Amira Abdel-Rahman, Nayeemur Rahman, Leandra Tejedor, Mackenzie Leake, Megan Hofmann, and Stefanie Mueller. 2023. Style2Fab: Functionality-aware segmentation for fabricating personalized 3d models with generative ai. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–13.
[19] Charles Finance and Susan Zwerman. 2015. *The visual effects producer: understanding the art and business of VFX*. Routledge.
[20] Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. 2024. Scene-llm: Extending language model for 3d visual understanding and reasoning. *arXiv preprint arXiv:2403.11401* (2024).
[21] Gege Gao, Weiyang Liu, Anpei Chen, Andreas Geiger, and Bernhard Schölkopf. 2024. Graphdreamer: Compositional 3d scene synthesis from scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21295–21304.
[22] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. 2022. Get3d: A generative model of high quality 3d textured shapes learned from images. *Advances in neural information processing systems* 35 (2022), 31841–31854.
[23] Zeqi Gu, Yin Cui, Zhaoshuo Li, Fangyin Wei, Yunhao Ge, Jinwei Gu, Ming-Yu Liu, Abe Davis, and Yifan Ding. 2025. ArtiScene: Language-Driven Artistic 3D Scene Generation Through Image Intermediary. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2891–2901.
[24] Jiaqi Guo, Lianli Gao, Junchen Zhu, Jiaxin Zhang, Siyang Li, and Jingkuan Song. 2024. MagicVFX: Visual Effects Synthesis in Just Minutes. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 8238–8246.
[25] Erin J Hastings, Ratan K Guha, and Kenneth O Stanley. 2008. Interactive evolution of particle systems for computer graphics and animation. *IEEE Transactions on Evolutionary Computation* 13, 2 (2008), 418–432.
[26] Hao-Yu Hsu, Chih-Hao Lin, Albert J Zhai, Hongchi Xia, and Shenlong Wang. 2025. Autovfx: Physically realistic video editing from natural language instructions. In *2025 International Conference on 3D Vision (3DV)*. IEEE, 769–780.
[27] Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, et al. 2024. Chat-scene: Bridging 3d scene and large language models with object identifiers. *Advances in Neural Information Processing Systems* 37 (2024), 113991–114017.
[28] Ian Huang, Yanan Bao, Karen Truong, Howard Zhou, Cordelia Schmid, Leonidas Guibas, and Alireza Fathi. 2025. Fireplace: Geometric refinements of llm common sense reasoning for 3d object placement. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 13466–13476.
[29] Nitish Joshi, Abulhair Saparov, Yixin Wang, and He He. 2024. Llms are prone to fallacies in causal inference. *arXiv preprint arXiv:2406.12158* (2024).
[30] Joonho Kim, Fanny Chevalier, and Karan Singh. 2025. Squidgets: Sketch-based Widget Design for Scene Manipulation. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*. 1–13.
[31] Mert Kiray, Paul Uhlenbruck, Nassir Navab, and Benjamin Busam. 2025. PromptVFX: Text-Driven Fields for Open-World 3D Gaussian Animation. *arXiv preprint arXiv:2506.01091* (2025).
[32] Dohwan Ko, Ji Soo Lee, Wooyoung Kang, Byungseok Roh, and Hyunwoo J Kim. 2023. Large language models are temporal and causal reasoners for video question answering. *arXiv preprint arXiv:2310.15747* (2023).
[33] Haoran Li, Haolin Shi, Wenli Zhang, Wenjun Wu, Yong Liao, Lin Wang, Lik-hang Lee, and Peng Yuan Zhou. 2024. Dreamscene: 3d gaussian-based text-to-3d scene generation via formation pattern sampling. In *European Conference on Computer Vision*. Springer, 214–230.
[34] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–15.
[35] Lu Ling, Chen-Hsuan Lin, Tsung-Yi Lin, Yifan Ding, Yu Zeng, Yichen Sheng, Yunhao Ge, Ming-Yu Liu, Aniket Bera, and Zhaoshuo Li. 2025. Scenethesis: A language and vision agentic framework for 3d scene generation. *arXiv preprint arXiv:2505.02836* (2025).
[36] Qihao Liu, Yi Zhang, Song Bai, Adam Kortylewski, and Alan Yuille. 2024. Direct-3d: Learning direct text-to-3d generation on massive noisy 3d data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6881–6891.
[37] Xinyu Liu, Ailing Zeng, Wei Xue, Harry Yang, Wenhan Luo, Qifeng Liu, and Yike Guo. 2025. VFX Creator: Animated Visual Effect Generation with Controllable Diffusion Transformer. *arXiv preprint arXiv:2502.05979* (2025).
[38] Tom Livingstone. 2024. Game engines: Optimising VFX, reshaping visual media. (2024).
[39] Sebastian Misztal and Jonas Schild. 2022. Visual delegate generalization frame–evaluating impact of visual effects and elements on player and user experiences in video games and interactive virtual environments. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–20.
[40] Başak Melis Öcal, Maxim Tatarchenko, Sezer Karaoğlu, and Theo Gevers. 2024. Sceneteller: Language-to-3d scene generation. In *European Conference on Computer Vision*. Springer, 362–378.
[41] Jeongseok Oh and SeungJun Kim. 2025. MoWa: An Authoring Tool for Refining AI-Generated Human Avatar Motions Through Latent Waveform Manipulation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–21.
[42] Jeongseok Oh, Seungju Kim, and Seungjun Kim. 2024. LumiMood: A creativity support tool for designing the mood of a 3D scene. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–21.
[43] Jeffrey A Okun, VES Susan Zwerman, et al. 2020. *The VES handbook of visual effects: industry standard VFX practices and procedures*. Routledge.
[44] Haiyi Qiu, Minghe Gao, Long Qian, Kaihang Pan, Qifan Yu, Juncheng Li, Wenjie Wang, Siliang Tang, Yueting Zhuang, and Tat-Seng Chua. 2025. STEP: Enhancing Video-LLMs' Compositional Reasoning by Spatio-Temporal Graph-guided Self-Training. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 3284–3294.
[45] William T Reeves. 1998. Particle systems—a technique for modeling a class of fuzzy objects. In *Seminal graphics: pioneering efforts that shaped the field*. 203–220.
[46] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).
[47] Giovanni Ribeiro, Katja Rogers, Maximilian Altmeyer, Thomas Terkildsen, and Lennart E Nacke. 2020. Game atmosphere: effects of audiovisual thematic cohesion on player experience and psychophysiology. In *Proceedings of the annual symposium on computer-human interaction in play*. 107–119.
[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.
[49] Vishnu Sarukkai, Lu Yuan, Mia Tang, Maneesh Agrawala, and Kayvon Fatahalian. 2024. Block and detail: Scaffolding sketch-to-image generation. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*.

1–13.

[50] Francisco G Sepúlveda. 2020. *Visual aesthetics in video games and their effects on player motivation.* Northeastern University.

[51] Xinyu Shi, Li-Yi Wei, Nanxuan Zhao, Jian Zhao, and Rubaiat Habib Kazi. 2026. Notational animating: An interactive approach to creating and editing animation keyframes. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–27.

[52] Ben Shneiderman. 2007. Creativity support tools: accelerating discovery and innovation. *Commun. ACM* 50, 12 (2007), 20–32.

[53] Yawar Siddiqui, Tom Monnier, Filippos Kokkinos, Mahendra Kariya, Yanir Kleiman, Emilien Garreau, Oran Gafni, Natalia Neverova, Andrea Vedaldi, Roman Shapovalov, et al. 2024. Meta 3d assetgen: Text-to-mesh generation with high-quality geometry, texture, and pbr materials. *Advances in Neural Information Processing Systems* 37 (2024), 9532–9564.

[54] Aswathy Sreenivasan and M Suresh. 2024. Design thinking and artificial intelligence: A systematic literature review exploring synergies. *International Journal of Innovation Studies* 8, 3 (2024), 297–312.

[55] Unity Technologies. [n. d.]. Particle effects. https://docs.unity3d.com/6000.2/Documentation/Manual/ParticleSystems.html

[56] Christina Tsalicoglou, Fabian Manhardt, Alessio Tonioni, Michael Niemeyer, and Federico Tombari. 2023. Textmesh: Generation of realistic 3d meshes from text prompts. *arXiv preprint arXiv:2304.12439* (2023).

[57] Veera Vimpari, Annakaisa Kultima, Perttu Hämäläinen, and Christian Guckelsberger. 2023. "An adapt-or-die type of situation": perception, adoption, and use of text-to-image-generation AI by game industry professionals. *Proceedings of the ACM on Human-Computer Interaction* 7, CHI PLAY (2023), 131–164.

[58] Jianren Wang and Yihui He. 2019. Physics-aware 3d mesh synthesis. In *2019 International Conference on 3D Vision (3DV)*. IEEE, 502–512.

[59] Xingkui Wei, Zhengqing Chen, Yanwei Fu, Zhaopeng Cui, and Yinda Zhang. 2021. Deep hybrid self-prior for full 3D mesh generation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5805–5814.

[60] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. 2025. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 21469–21480.

[61] Jun Xing, Rubaiat Habib Kazi, Tovi Grossman, Li-Yi Wei, Jos Stam, and George Fitzmaurice. 2016. Energy-brushes: Interactive tools for illustrating stylized elemental dynamics. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. 755–766.

[62] Yiying Yang, Fukun Yin, Jiayuan Fan, Xin Chen, Wanzhang Li, and Gang Yu. 2024. Scene123: One prompt to 3d scene generation via video-assisted and consistency-enhanced mae. *arXiv preprint arXiv:2408.05477* (2024).

[63] Georgios N Yannakakis, Antonios Liapis, and Constantine Alexopoulos. 2014. Mixed-initiative co-creativity. (2014).

[64] Hang Yin, Xiuwei Xu, Zhenyu Wu, Jie Zhou, and Jiwen Lu. 2024. Sg-nav: Online 3d scene graph prompting for llm-based zero-shot object navigation. *Advances in neural information processing systems* 37 (2024), 5285–5307.

[65] Yuqian Yuan, Hang Zhang, Wentong Li, Zesen Cheng, Boqiang Zhang, Long Li, Xin Li, Deli Zhao, Wenqiao Zhang, Yueting Zhuang, et al. 2025. Videorefer suite: Advancing spatial-temporal object understanding with video llm. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 18970–18980.

[66] Bingqing Zhang and Wenfeng Hu. 2017. Game special effect simulation based on particle system of Unity3D. In *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*. IEEE, 595–598.

[67] Lei Zhang, Jin Pan, Jacob Gettig, Steve Oney, and Anhong Guo. 2024. VRCopilot: authoring 3D layouts with generative AI models in VR. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–13.

[68] Yifei Zhang, Lin-Ping Yuan, Yuheng Zhao, Jielin Feng, and Siming Chen. 2025. KinemaFX: A Kinematic-Driven Interactive System for Particle Effect Exploration and Customization. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*. 1–17.

[69] Eric Zhou and Dokyun Lee. 2024. Generative artificial intelligence, human creativity, and art. *PNAS nexus* 3, 3 (2024), pgae052.

A Comparison Framework Formulations

A.1 Task 1: Ablated Framework

Without prompt enhancement, $\mathcal{L}_{\text{proj}}$ and $\mathcal{L}_{\text{causality}}$ are unavailable. The simplified losses become:

$$\mathcal{L}_{\text{spatial}}^{\text{ablated}} = \mathcal{L}_{\text{init}} + \mathcal{L}_{\text{database}}, \quad \mathcal{L}_{\text{temporal}}^{\text{ablated}} = \mathcal{L}_{\text{init}} + \mathcal{L}_{\text{coverage}} + \mathcal{L}_{\text{database}} \tag{8}$$

All individual terms are defined in Sections 4.2 and 4.3.

A.2 Task 2: 3D Generation Framework

With Trellis replacing the design modules, initial layouts are derived from prompt enhancement: spatial positions from 2D bounding box centers mapped to 3D, temporal delays from causality order:

$$\mathbf{p}_i^{\text{init}} = \begin{bmatrix} (x_{\text{center}} - 0.5) \cdot s \\ (y_{\text{center}} - 0.5) \cdot s \\ z_{\text{default}} \end{bmatrix}, \quad t_i^{\text{init}} = \mathit{causality_order}(c_i) \cdot \Delta t \tag{9}$$

The losses retain SpeLL's full structure with $\mathcal{L}_{\text{init}}$ anchored to these prompt-derived values:

$$\mathcal{L}_{\text{spatial}}^{\text{3DGen}} = \mathcal{L}_{\text{init}} + \mathcal{L}_{\text{proj}} + \mathcal{L}_{\text{database}}, \quad \mathcal{L}_{\text{temporal}}^{\text{3DGen}} = \mathcal{L}_{\text{init}} + \mathcal{L}_{\text{causality}} + \mathcal{L}_{\text{database}} + \mathcal{L}_{\text{coverage}} \tag{10}$$

B Questionnaires

Table 4: Task 1 evaluation questionnaire measuring user experience and system performance.

Dimension	Question (Likert 1-7)
Intention Alignment	The AI-generated VFX aligns with my intended design.
Interpretation Fidelity	The AI accurately understood my prompt.
Outcome Aesthetics	The quality of the AI-generated VFX is high.
Spatial Layout	The spatial arrangement of VFX components is satisfactory.
Temporal Layout	The timing of component appearances and disappearances is satisfactory.
Perceived Utility	I can create high-quality VFX based on this AI output.

Table 5: Task 2 evaluation questionnaire comparing generation models for VFX production.

Dimension	Question (Likert 1-7)
Outcome Quality	The overall quality of the AI-generated VFX is high
Topology Suitability	The generated mesh topology is highly suitable for VFX production
Texture Suitability	The applied textures are highly appropriate for VFX integration
Editability	The generated results appear easy to modify as desired
Production Viability	I can create high-quality VFX based on this AI output
Versatility	The AI results facilitate creation of diverse variations

Table 6: Task 3 Creativity Support Index questionnaire.

Dimension	Questions (Likert 1-10)
Enjoyment	I would be happy to use this system on a regular basis. I enjoyed using the system or tool.
Exploration	I was able to explore various ideas, design options, and outcomes using this approach. This approach helps in finding diverse ideas and results.
Expressiveness	I was able to be very creative while doing the activity inside this system or tool. The system or tool allowed me to be very expressive.
Immersion	My attention was fully tuned to the activity, and I forgot about the system that I was using. I became so absorbed in the activity that I forgot about the system that I was using.
Results Worth Effort	I was satisfied with what I got out of the system. What I was able to produce was worth the effort I had to exert to produce it.

C Statistics

Table 7: Statistical results for all three tasks showing means, standard deviations, and significance tests.

Task	Statistics & Significance Test			Statistics & Significance Test		
Task 1 - Questionnaire	**Intention Alignment**			**Interpretation Fidelity**		
	Ablated	$M = 1.60$, $SD = 0.82$	$W = 0.00$	*Ablated*	$M = 2.00$, $SD = 1.38$	$W = 12.50$
	SpeLL	$M = 4.15$, $SD = 1.76$	$p < .001^{***}$	*SpeLL*	$M = 4.05$, $SD = 1.61$	$p < .001^{***}$
	Outcome Aesthetics			**Spatial Layout**		
	Ablated	$M = 2.20$, $SD = 1.36$	$t(19) = -4.72$	*Ablated*	$M = 2.75$, $SD = 1.41$	$t(19) = -4.265$
	SpeLL	$M = 4.00$, $SD = 1.81$	$p < .001^{***}$	*SpeLL*	$M = 4.25$, $SD = 1.12$	$p < .001^{***}$
	Temporal Layout			**Perceived Utility**		
	Ablated	$M = 3.20$, $SD = 1.77$	$t(19) = -3.09$	*Ablated*	$M = 2.40$, $SD = 1.64$	$t(19) = -6.10$
	SpeLL	$M = 4.55$, $SD = 1.70$	$p = .006^{**}$	*SpeLL*	$M = 4.45$, $SD = 1.85$	$p < .001^{***}$
Task 1 - Prompts	**RM ANOVA**	**Post-hoc Test**				
	$F(1.23, 23.39) = 7.638$	*Effect - Component*	$p = .004^{**}$			
	$p = .008^{**}$	*Effect - Layout*	$p = .006^{**}$			
		Component - Layout	$p = 1.000$			
Task 2 - Questionnaire	**Outcome Quality**			**Topology Suitability**		
	3DGen	$M = 2.20$, $SD = 1.06$	$W = 7.00$	*3DGen*	$M = 2.15$, $SD = 1.35$	$t(19) = -4.03$
	SpeLL	$M = 3.90$, $SD = 1.86$	$p = .002^{**}$	*SpeLL*	$M = 3.50$, $SD = 1.64$	$p < .001^{***}$
	Texture Suitability			**Editability**		
	3DGen	$M = 2.15$, $SD = 1.31$	$t(19) = -3.76$	*3DGen*	$M = 3.25$, $SD = 2.00$	$t(19) = -4.50$
	SpeLL	$M = 3.55$, $SD = 1.85$	$p = .001^{**}$	*SpeLL*	$M = 5.40$, $SD = 1.40$	$p < .001^{***}$
	Production Viability			**Versatility**		
	3DGen	$M = 2.00$, $SD = 1.49$	$t(19) = -3.54$	*3DGen*	$M = 2.40$, $SD = 1.50$	$W = 10.00$
	SpeLL	$M = 3.55$, $SD = 2.01$	$p = .002^{**}$	*SpeLL*	$M = 3.75$, $SD = 2.07$	$p = .002^{**}$
Task 3 - Questionnaire	**Enjoyment**			**Exploration**		
	Manual	$M = 3.75$, $SD = 2.09$	$t(19) = -3.71$	*Manual*	$M = 4.45$, $SD = 2.68$	$t(19) = -3.05$
	SpeLL	$M = 6.20$, $SD = 2.23$	$p = .001^{**}$	*SpeLL*	$M = 6.25$, $SD = 2.26$	$p = .007^{**}$
	Expressiveness			**Immersion**		
	Manual	$M = 4.70$, $SD = 2.66$	$t(19) = -1.82$	*Manual*	$M = 4.80$, $SD = 2.36$	$W = 34.50$
	SpeLL	$M = 5.65$, $SD = 2.21$	$p = .084$	*SpeLL*	$M = 5.83$, $SD = 2.52$	$p = .048^{*}$
	Result Worth Effort					
	Manual	$M = 3.38$, $SD = 2.15$	$t(19) = -3.74$			
	SpeLL	$M = 5.83$, $SD = 2.66$	$p = .001^{**}$			
Task 3 - Quantitative	**Task Completion Time**			**Interaction Count**		
	Manual	$M = 4.77$, $SD = 4.93$	$W = 15.00$	Manual	$M = 129.50$, $SD = 136.23$	$t(19) = -3.51$
	SpeLL	$M = 1.16$, $SD = 1.59$	$p = .02^{*}$	SpeLL	$M = 61.70$, $SD = 74.19$	$p = .002^{**}$