Dynamic Diorama: Narrative-Driven Orientation Modeling and Object Placement for VR

Furkan Çelen
Computer Engineering Dept.
Istanbul Technical University
Istanbul, Türkiye
email: celen23f@itu.edu.tr

Meral Kuyucu
Computer Engineering Dept.
Istanbul Technical University
Istanbul, Türkiye
email: korkmazmer@itu.edu.tr

Bora Şenceylan
Computer Engineering Dept.
Istanbul Technical University
Istanbul, Türkiye
email: senceylan19@itu.edu.tr

Gökhan İnce
Computer Engineering Dept.
Istanbul Technical University
Istanbul, Türkiye
email: gokhan.ince@itu.edu.tr

***Abstract*—Spatial composition is a key factor in Virtual Reality storytelling, as object arrangement directly influences how users perceive meaning and emotion. However, converting text into 3D layouts is difficult because systems typically prioritize geometric rules over narrative context. Dynamic Diorama moves beyond simple geometry by analyzing text for structural details and emotional cues to shape the layout. Rather than just placing objects randomly or by rigid rules, our pipeline aligns the spatial relationships directly with the story's mood. This approach is benchmarked against standard baselines across four distinct narrative themes: happiness, fear, surprise, and sadness. Participant feedback indicated that the Large Language Model driven scenes offered significantly better narrative coherence and emotional alignment compared to the random and heuristic baselines. Eye-tracking data supported this finding, revealing that semantically informed scenes reduced the time-to-first-fixation on key narrative elements.**

***Keywords*—narrative-driven scene generation; VR storytelling; semantic layout reasoning; spatial placement.**

I. Introduction

Stories in Virtual Reality (VR) are shaped as much by spatial arrangement as by plot. Where an object sits, how characters face each other, and which items occupy the foreground all contribute to what users notice and how they feel. These decisions are not merely authoring challenges, but perceptual ones, as spatial decisions directly shape what users attend to and how narratives are experienced. Unlike traditional screen-based media, immersive Augmented Reality (AR) and VR environments amplify the perceptual impact of spatial decisions. Object placement, orientation, and proximity are not merely compositional choices but directly influence users' sense of presence, attention, and emotional engagement. Small spatial inconsistencies may disrupt narrative flow or reduce immersion, while subtle spatial cues can silently reinforce tension, intimacy, or anticipation.

Creating spatial narratives remains labor-intensive, as designers must constantly trade off physical plausibility, visual clarity, and affective intent—factors that do not always align. Many automated layout tools prioritize geometric or visibility constraints, producing scenes that are technically coherent but may fail to reflect the intended emotional cues of the story [1][2]. Narrative text contains both explicit relations (e.g., "the lamp sits on the table") and implicit affective cues, such as tension, intimacy, or distance, that can inform scene composition. Still, prior evaluations tend to emphasize per-object plausibility or geometry-focused metrics, with comparatively less attention given to human perception of narrative coherence in immersive environments. Recent language-conditioned systems demonstrate that free-text can guide object placement and reduce manual effort [3][4].

Advances in Artificial Intelligence (AI), particularly language-based models, offer an opportunity to bridge this gap between narrative intent and spatial realization. Narrative descriptions often encode affective and relational cues implicitly rather than explicitly, leaving room for interpretation that exceeds the expressiveness of rule-based systems. In this context, AI-driven reasoning is not introduced to maximize automation, but to explore whether semantic and affect-aware interpretations of narrative text can better support spatial storytelling in immersive environments. To study this interpretive space systematically, this study frames placement as an experimental variable of narrative-driven scene generation.

We introduce a controlled VR testbed, Dynamic Diorama, which exposes three placement paradigms (random baseline, rule-based heuristic, and Large Language Model (LLM) informed placement) and applies them to short vignettes designed to evoke distinct emotions. The testbed combines semantic parsing, heuristic validation, and placement ranking to maintain physical plausibility while enabling systematic comparisons across placement approaches. We investigated how variations in placement strategy shape viewer perception, attention, and spatial validity. This study addresses the following research questions:

- **RQ1:** How do layouts produced by different object placement strategies affect viewers' perceived narrative coherence and emotional alignment in VR scenes?
- **RQ2:** How do these layout strategies influence visual attention patterns, as measured through gaze-based metrics, during narrative scene exploration?
- **RQ3:** To what extent can layout strategies that incorporate higher-level semantic or affective reasoning improve perceived narrative quality without increasing spatial invalidity, such as collisions or physically implausible placements?

The contributions of this study are threefold: 1) a controlled experimental testbed for narrative scene layout in VR, where the object placement strategy is treated as an independent experimental variable, 2) an experimental setup with three random, heuristic, and LLM-based layout strategies designed to enable systematic comparison of their effects on viewer perception and attention and 3) an empirical VR study that examines the effects of different layout strategies on narrative coherence, emotional alignment, and visual attention using self-report and gaze-based measures.

The remainder of this paper is organized as follows: Section II reviews related work on text-to-scene generation. Section III details the Dynamic Diorama framework and placement strategies. Section IV describes the experimental design and methodology. Section V presents the quantitative and qualitative results. Finally, Section VI concludes the study and outlines future directions.

II. Related Work

A. Object Placement in Text-to-Scene Systems

Early attempts to generate scenes from text relied heavily on hand-crafted grammars and domain-specific rules, which proved brittle when applied to open-ended or creative narratives [2]. Recent work has shifted toward transformer-based language models, which are better suited to handling ambiguity and implicit structure. LLMs have been shown to extract elements, such as scene boundaries, characters, and relationships directly from unstructured natural language [5]. Systems like PlaceIt3D [3] and SceneTeller [4] illustrate how natural language can be used to guide 3D layout generation and reduce the need for manual scene construction.

In VR and Mixed Reality (MR) applications, object placement is typically constrained by geometric and semantic considerations. Common practices include managing occlusion, optimizing visibility, and respecting surface affordance, such as ensuring that objects are placed on appropriate supports rather than floating in space [1][3]. Work on AR label placement and occlusion-aware heuristics highlights the limits of geometry-first optimization when communicative or narrative goals are considered [6][7][8]. Scene-graph representations further formalize spatial relationships by linking object categories to likely locations and neighboring elements, demonstrating how semantic information supports functional placement [9]. These methods are effective at producing physically consistent environments.

Still, many of these approaches focus on identifying what should appear in a scene, paying less attention to how the emotional tone of a narrative should influence spatial composition. However, physical correctness alone does not guarantee that a scene supports a narrative. In many cases, layouts that are spatially sound fail to convey the emotional tension or intimacy implied by a story, resulting in environments that feel correct but emotionally unengaging [6].

B. Learning-Based Object Placement

Learning-based object placement models address some of the limitations of rule-driven systems by inferring spatial patterns from large collections of scenes [10]. This allows them to capture contextual relationships that are difficult to express through explicit heuristics. At the same time, unconstrained learning-based outputs can introduce practical issues, including object collisions or violations of physical affordance. Hybrid approaches attempt to balance these trade-offs by combining learned or language-derived proposals with rule-based validation mechanisms [11]. Such combinations are especially relevant in storytelling contexts, where a degree of spatial flexibility is needed, but basic physical coherence must still be preserved [12].

C. Interactive Layout Tools

Another line of work extends language-conditioned placement through open-vocabulary mappings, allowing free-form textual descriptions to be associated with 3D assets beyond fixed label sets [13]. This capability is particularly important for narrative scenes, where descriptions are often abstract or metaphorical. Interactive, chat-driven layout tools further point toward more fluid human–model workflows by enabling authors to iteratively refine layouts through dialogue rather than low-level parameter tuning, as demonstrated by systems, such as Chat2Layout [12]. Despite these advances, evaluating whether a generated scene actually aligns with an author's intent remains difficult. As a result, assessment often depends on a combination of automated measures and user-centered evaluation.

D. Emotion-Aware Spatial Design

Emotion-aware techniques are widely used in VR to drive reactive elements, such as lighting, sound, or avatar behavior [14]. Their influence on the spatial arrangement of objects, however, has received comparatively less attention. Frameworks like UniEmoX [15] suggest ways to model emotion perception in a general form, but there is still limited empirical evidence on how emotion-driven spatial composition affects user experience in immersive environments. This gap motivates our study that compares different placement strategies—ranging from random and heuristic methods to language-guided approaches—under controlled emotional narratives.

While existing text-to-scene systems primarily focus on geometric plausibility, our Dynamic Diorama framework introduces a novel approach by treating spatial placement as an experimental variable driven by affective cues. This explicitly bridges the gap between semantic LLM reasoning and emotional narrative alignment in VR.

III. Dynamic Diorama Framework

A. Framework Overview

In this study, we propose Dynamic Diorama as a research platform that supports systematic observation of how different object placement strategies influence narrative experience in

VR. This framework emphasizes comparability across conditions by using the same stories, assets, and physical constraints. This keeps the focus on placement behavior rather than content differences.

Figure 1 illustrates the high-level workflow of the Dynamic Diorama framework. In this workflow, the narrative text is first organized into scene-level representations that guide object placement. The resulting layouts are validated for spatial plausibility before being rendered as a VR story.

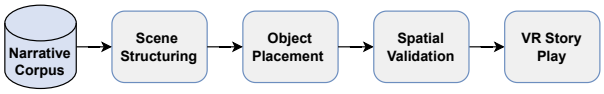


Figure 1. High-level workflow of the Dynamic Diorama framework.

B. Narrative Stimuli and Visual Scope

Four narratives were prepared to be used across all experiments to avoid variation. Each narrative was written to convey a single dominant emotion (happy, fearful, surprised, or sad). Narrative length ranged from approximately 60 to 100 words. To support temporal progression, each story is divided into five sequential scenes. Scene boundaries are derived using an LLM that produces a structured representation of narrative flow. Rather than generating geometry directly, this representation is later interpreted by the Unity runtime to drive scene transitions and placement logic. This step enables consistent story structure across placement strategies while keeping narrative content fixed.

All scenes are composed of a fixed pool of 30 3D assets. This collection comprises a diverse set of low-poly models, including environmental elements (e.g., furniture, foliage), human characters, and animals. These assets were selected for their narrative versatility, allowing the same objects to be recontextualized across different emotional scenarios (e.g., a dog functioning as a companion in a 'Happy' scene or a threat in a 'Fear' scene). The asset pool size was limited by hardware performance, development effort, and the requirements of standalone VR deployment. No placement approach is given access to additional assets or visual effects. Assets were reused across scenes, with differences arising only from their spatial arrangement, orientation, and relationships. Keeping the asset set fixed reduces the influence of visual richness and keeps spatial composition as the main experimental variable.

C. Scene Structuring

A narrative, N, can be represented as the following ordered sequence of scenes

$$N = \{s_1, s_2, \dots s_t \dots s_T\}, \tag{1}$$

where each scene s_t corresponds to a distinct segment of the story. For a given scene, the system operates over a fixed asset pool A, consisting of arbitrary assets (a_k) and is identical across all experimental conditions, as follows:

$$A = \{a_1, a_2, \dots a_k \dots a_K\}, \tag{2}$$

The outcome of scene composition is a spatial layout

$$L_t = \{(a_i, \mathbf{p}_i, \theta_i) \mid a_i \in A_t \subseteq A\}, \tag{3}$$

where $\mathbf{p}_i \in \mathbb{R}^3$ denotes the 3D position vector of asset a_i, and $\theta_i \in [0, 2\pi)$ represents its orientation around the vertical axis. Differences between placement approaches arise from how these layouts are produced.

The placement process is defined as mapping a scene description and asset pool to a spatial layout:

$$f : (s_t, A) \rightarrow L_t, \tag{4}$$

In this study, all placement approaches implement the same mapping interface f, but differ in the information used to guide it as: random placement ignores narrative semantics, heuristic-based placement relies on predefined spatial rules, and the LLM placement incorporates narrative and affective cues extracted from the text.

D. Object Placement

Three placement approaches are implemented within the same framework, each operating on identical narrative input and spatial limits but differing in how placement decisions are produced.

1) Constrained Random Placement: The random baseline is intentionally simple, while still enforcing spatial constraints. Object locations are generated without regard to narrative meaning or emotional tone. Basic checks are applied to avoid collisions and invalid placements. This ensures that scenes remain navigable and visually acceptable, despite lacking semantic or emotion-aware structure. This condition serves as a baseline reference for evaluation. The underlying logic for this approach is detailed in Figure 2.

Algorithm 1 Random Placement Strategy

Require: Asset pool A, valid spatial regions R
Ensure: Scene layout L
1: **for** each asset $a \in A$ **do**
2: Random position $p \sim R$
3: Random orientation $\theta \sim [0, 2\pi)$
4: Add (a, p, θ) to L
5: **end for**
6: **return** L

Figure 2. Pseudo-code for the random placement strategy.

2) Heuristic-based Placement: In the heuristic condition, object placement follows a fixed set of designer-defined rules specified in a JSON configuration file. These rules determine which assets may appear in each scene and constrain their allowable regions, orientations, and basic spatial relationships. The rules reflect common assumptions about plausible object

arrangements but are not sensitive to narrative emotion. As a result, the layouts are consistent and easy to interpret, but they do not change in response to the emotional tone of the story. This limitation is clearly illustrated in Figure 5, where the heuristic agent correctly orients towards the target but fails to exhibit the bodily expression required for the 'Surprised' emotion. In practice, the heuristics are implemented through spatial constraints, such as raycast-based surface checks, predefined anchor regions, and simple orientation rules. The execution flow of these rules is outlined in Figure 3.

Algorithm 2 Heuristic-Based Placement Strategy

Require: Asset pool A, rule set $\mathcal{H}$
Ensure: Scene layout L
1: **for** each asset $a \in A$ **do**
2: Retrieve rule $h \in \mathcal{H}$ corresponding to asset type of a
3: Determine position $p \leftarrow h_{\text{pos}}(a)$
4: Determine orientation $\theta \leftarrow h_{\text{ori}}(a)$
5: Add (a, p, θ) to L
6: **end for**
7: **return** L

Figure 3. Pseudo-code for the heuristic-based placement strategy.

3) LLM-Informed Placement: In the LLM-informed condition, the narrative text is interpreted by Google's Gemini 3 Pro model. We selected this model for its advanced reasoning capabilities in spatial context understanding and multimodal processing [16]. The model was accessed via API with a temperature setting of 0.7 to balance structural adherence with creative interpretation. It produces a structured description used by the Unity pipeline. This description encodes relational cues, such as relative distance, spatial priority, and orientation (e.g., near a window, facing the viewer), which are then translated into concrete 3D placements by the engine. Unlike the other placement approaches, this condition allows cues from the narrative to influence spatial decisions. These cues can affect object proximity, character orientation, facial expressions, and gaze direction, including how elements are oriented relative to the viewer. To maintain a clear separation between conditions, affect-driven adjustments are applied only in the LLM-informed placement strategy. For example, the structural adjustments made to visually emphasize the antagonist in the 'Fear' scenario are demonstrated in Figure 6. The LLM pipeline relies on a zero-shot prompting strategy where the scene boundaries and object relations are parsed into json objects. These objects dictate the exact spatial parameters (e.g., proxemics, gaze vectors) before the Unity engine renders the final layout. The integration of this semantic parsing into the pipeline is summarized in Figure 4.

E. Spatial Validation

All scenes undergo the same validation checks before rendering, regardless of placement strategy. These checks include collision detection, surface support verification, and spacing constraints. Applying the same validation across conditions prevents physically implausible layouts while still allowing differences in spatial composition. Validation is implemented using Unity's PhysX engine through collider intersection checks and surface support tests.

Algorithm 3 LLM-Informed Placement Strategy

Require: Narrative text N, asset pool A
Ensure: Relational placement specification R
1: Extract spatial relations $S \leftarrow \text{Analyze}(N)$
2: **for** each relation $r \in S$ **do**
3: Infer spatial parameters (distance, orientation, salience) for r
4: Associate parameters with relevant assets in A
5: **end for**
6: **return** R

Figure 4. Pseudo-code for the LLM-Informed placement strategy.

IV. Evaluation

A. Hardware and Software

The study was conducted using the HTC Vive Focus Vision headset. All scenes were rendered in real time using Unity 6, and the same application build was used across all experiments. Audio was delivered through the headset's built-in speakers.

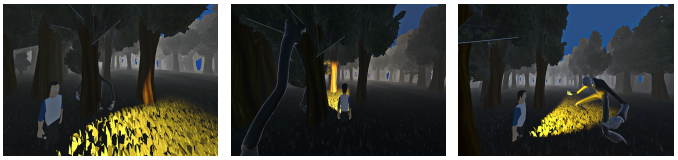
Interaction was limited to natural head movement, keeping attention on the narrative and the surrounding scene rather than on controls. All sessions were conducted in the same physical environment, with room layout, ambient lighting, and verbal instructions kept consistent to minimize external variation. To capture granular attention metrics, we leveraged the HTC Vive Focus Vision's built-in eye-tracking capabilities via the OpenXR interface. Gaze origin and direction vectors were accessed in real-time, synchronized with the application's update loop. We implemented a custom raycasting system that continuously mapped these vectors to the 3D scene geometry, allowing the system to log specific object fixations, gaze duration, and scan paths with high precision throughout the narrative experience [17].

B. Experimental Design

Participants joined the experiment voluntarily and were recruited informally. No specific experience with VR, games, or interactive storytelling was required. Each participant completed a single session in which the same narrative was presented three times under different scene configurations. The narrative text, asset pool, and runtime constraints remained unchanged, so differences between scene versions were limited to object placement and spatial relationships. Scene configurations were labeled neutrally (Scene A, B, and C), and participants were not informed of the underlying placement strategies to avoid expectation bias. The order of the three scene versions was random between participants to reduce sequence effects.

(a) Random-Based (b) Heuristic-Based (c) LLM-Based

Figure 5. 'Surprised' scenario (Diorama 1). (a) **Random-Based:** Character gazes at an arbitrary point, lacking context. (b) **Heuristic-Based:** Character correctly looks at the target (bird) based on rules, but the body orientation fails to convey the emotional state to the viewer. (c) **LLM-Based:** Character maintains gaze on the target while orienting the body towards the camera, effectively displaying the 'Surprised' emotion and maximizing user immersion.



(a) Random-Based (b) Heuristic-Based (c) LLM-Based

Figure 6. 'Fear' scenario (Diorama 2). (a) **Random-Based:** Character gazes aimlessly into the dark forest, failing to acknowledge the nearby antagonist. (b) **Heuristic-Based:** Character adheres to the gaze rule by facing the target coordinates, but the composition fails due to occlusion, leaving the antagonist visually obstructed by a tree. (c) **LLM-Based:** The scene is semantically restructured to reveal the antagonist clearly, while the character's body orientation and gaze align to vividly portray the 'Fear' state to the viewer.

(a) Random-Based (b) Heuristic-Based (c) LLM-Based

Figure 7. 'Sadness' scenario (Diorama 3). (a) **Random-Based:** Character stands in the background facing away from the focal point (photograph), completely missing the narrative beat. (b) **Heuristic-Based:** Character satisfies the gaze constraint by facing the target, but the substantial distance and rigid posture fail to evoke the intended intimate atmosphere. (c) **LLM-Based:** The agent demonstrates semantic understanding of proxemics by positioning the character intimately close to the memento and adjusting the head tilt downward, effectively embodying the emotion of mourning through non-verbal cues.

(a) Random-Based (b) Heuristic-Based (c) LLM-Based

Figure 8. 'Happiness' scenario (Diorama 4). (a) **Random-Based:** Spatial layout is disjointed; the secondary character (dog) is obscured by foliage in the background, severing the narrative link. (b) **Heuristic-Based:** The agent aligns orientation to the target, but the excessive physical distance results in a detached observation rather than a shared emotional experience. (c) **LLM-Based:** The model interprets 'Happiness' as active companionship, significantly reducing the spatial distance to foster a sense of intimacy and interaction between the character and the animal.

Each session began with a brief verbal introduction. Participants were informed that they would experience a VR-based story and later answer questions about their impressions and observations. Each participant was assigned one of four narratives, each centered on a single dominant emotional tone. The selected story was presented three times, once for each scene configuration, with short pauses between viewings to allow participants to rest without disrupting the overall narrative context.

After each scene viewing, participants completed a short questionnaire assessing narrative coherence, emotional alignment, and presence for the scene they had just experienced. Narrative coherence was evaluated through questions focusing on the relationship between object placement and story understanding, specifically how logical the arrangement of objects felt within the context of the story and whether placement helped direct attention to important elements. Presence was assessed by asking participants whether they felt physically present in the environment, became engrossed in the virtual world, and felt like part of the story rather than an external observer. Emotional alignment was measured by asking participants to identify the dominant emotion they felt in the scene and to rate how strongly that emotion was conveyed. After all three scene versions had been viewed, participants completed an additional comparative questionnaire. These questions asked participants to directly select which scene best conveyed the story's emotion, supported narrative understanding, guided visual attention most naturally, elicited the strongest sense of immersion, and which version they would choose to show to another person.

C. Measures and Data Collection

A total of 20 participants (16 male, 4 female) ranging in age from 21 to 45 participated in the study. The cohort possessed diverse educational backgrounds, predominantly holding or pursuing undergraduate (50%) and graduate (45%) degrees. The majority of the participants (90%) reported having beginner-level experience with VR technologies, while 10% possessed intermediate familiarity. Participant background information also included field of study (e.g., Computer Engineering, Electronics) to represent a wider range of user perspectives.

Responses were collected using a combination of short-answer prompts and 7-point Likert-scale ratings. For constructs measured using multiple Likert-scale items, responses were averaged to form composite scores for analysis.

V. Results

A. Results on Narrative Perception

To address RQ1, we examined how different object placement strategies influenced participants' perceptions of narrative coherence, emotional alignment, and presence. A one-way Analysis of Variance (ANOVA) was conducted to evaluate the statistical significance of the perception scores. As shown in Table I, the *LLM-based* placement strategy consistently received higher ratings in semantic categories. It achieved significantly higher perceived narrative coherence ($\mu = 5.7, \sigma = 0.9$) and stronger emotional alignment ($\mu = 5.4, \sigma = 0.9$) compared to the heuristic-based and random conditions ($p < 0.001$).

Interestingly, the sense of presence was comparable between the *Heuristic* ($\mu = 5.1, \sigma = 0.8$) and *LLM-based*

($\mu = 5.0, \sigma = 0.8$) conditions ($p > 0.05$). This suggests that while rule-based layouts can achieve physical plausibility, semantic reasoning is essential for conveying the narrative's emotional context and improving the storytelling capability of the scene.

TABLE I
PERCEPTION SCORES ACROSS PLACEMENT APPROACHES (N=20).

Metric	Random	Heuristic	LLM-based
Narrative Coherence	3.0 ± 1.7	5.1 ± 1.3	$\mathbf{5.7 \pm 0.9}$
Emotional Alignment	3.2 ± 1.5	4.4 ± 1.3	$\mathbf{5.4 \pm 0.9}$
Sense of Presence	4.4 ± 1.1	$\mathbf{5.1 \pm 0.8}$	5.0 ± 0.8

B. Results on Visual Attention

To address RQ2, we examined how object placement strategies influenced visual attention during scene viewing. As shown in Table II, layouts generated using the LLM-informed strategy consistently guided attention more effectively than the other conditions. Participants oriented to relevant objects more quickly in the LLM-based scenes. This is reflected in a significantly shorter *Time to First Fixation* ($\mu = 2.1\,\text{s}, \sigma = 0.7$) compared to the heuristic-based ($\mu = 3.6\,\text{s}, \sigma = 1.0$) and random layouts ($\mu = 4.8\,\text{s}, \sigma = 1.2$). Once fixated, participants also spent more time attending to key story elements in the LLM-informed condition, exhibiting longer *Dwell Times* ($\mu = 5.4\,\text{s}$) than in the other two conditions.

TABLE II
GAZE-BASED ATTENTION RESULTS ACROSS PLACEMENT APPROACHES.

Metric	Random	Heuristic	LLM-based
Time to First Fixation (s)	4.8 ± 1.2	3.6 ± 1.0	$\mathbf{2.1 \pm 0.7}$
Dwell Time (s)	2.3 ± 0.9	3.1 ± 1.1	$\mathbf{5.4 \pm 1.3}$

The differences in both Time to First Fixation and Dwell Time across the three conditions were found to be statistically significant ($p < 0.01$) using a one-way ANOVA.

C. Results on Physical Plausibility

We analyzed the physical plausibility metrics summarized in Table III. The *Random* baseline exhibited the highest instability, with an 11.2% initial collision rate and 7 visible artifacts in the final scenes, demonstrating the necessity of constraints. The *Heuristic* approach remained the most stable (2.1% collision) due to rigid rules.

The *LLM-based* strategy showed a moderate initial collision rate (3.8%), primarily driven by the model's semantic attempts to create intimate object proximity. However, the Spatial Validation layer effectively mitigated these risks. Consequently, the LLM-based approach achieved a highly plausible final result with minimal artifacts (2 instances), significantly outperforming the Random baseline and approaching the stability of handcrafted heuristics.

Since the physical plausibility metrics primarily consist of frequency counts, a Chi-square test was utilized, confirming that the reduction in invalid placements compared to the random baseline was statistically significant ($p < 0.05$).

TABLE III
PHYSICAL VALIDITY AND SYSTEM PERFORMANCE METRICS.

Metric	Random	Heuristic	LLM-based
Initial Collision Rate (%)	11.2	2.1	3.8
Validation Rejection Rate (%)	14.5	1.5	2.4
Avg. Generation Retries	1.8	0.2	0.8
Final Invalid Placements (Count)	**7**	**1**	**2**

*Refers to minor artifacts visible across all 20 experimental sessions.

D. Qualitative Results

Participants were asked to select the diorama version they felt best represented the story. As summarized in Table IV, the majority of participants (60%) explicitly preferred the *LLM-based* layouts. Qualitative feedback indicated that users found these scenes "more alive" and "narratively accurate," particularly praising the meaningful interactions between characters, such as the intimate proximity to the dog in the 'Happiness' scenario in Figure 8 or the mourning posture in the 'Sadness' scenario in Figure 7.

The *Heuristic* approach was preferred by 35% of users, primarily for its "clean and organized" structure, though some described it as "emotionally distant." The *Random* baseline was largely rejected (5%), with participants citing that the chaotic placement often "broke the immersion" and made the story difficult to follow. These preference rates align with the quantitative gains in coherence and emotional alignment reported in RQ1, confirming that users prioritize semantic depth over simple geometric order.

TABLE IV
USER PREFERENCES ACROSS PLACEMENT APPROACHES (N=20).

Preferred Scene Version	Percentage (%)
Random	5.0
Heuristic	35.0
LLM-based	**60.0**

In summary, the qualitative feedback and user preferences strongly corroborate the quantitative gaze and perception metrics, confirming that semantic layout reasoning significantly enhances the immersive storytelling experience.

VI. CONCLUSION AND FUTURE WORK

This study presented a framework that reimagines spatial layout not merely as a geometric puzzle, but as a narrative medium. By incorporating LLMs into the VR pipeline, we explored how spatial composition can move beyond static rules to reflect emotional context—translating abstract sentiments like "happiness" or "fear" into concrete proximity and orientation adjustments.

Our evaluation suggests that semantically informed layouts offer a tangible advantage in narrative coherence and emotional alignment over traditional baselines. Although the generative approach introduced high initial collision rates due to its ambitious placement strategies, our findings confirm that a validation layer can effectively mitigate these risks, balancing semantic expressiveness with physical plausibility.

For future work, we plan to integrate procedural mesh deformation and inverse kinematics to resolve physical conflicts dynamically, ensuring that intimate character interactions remain both semantically powerful and geometrically seamless without relying on expensive regeneration cycles. The evaluations will involve a larger and more diverse participant sample to further validate these findings.

REFERENCES

[1] M. Billinghurst, A. Clark, and G. Lee, "A survey of augmented reality," *Foundations and Trends in Human-Computer Interaction*, vol. 8, no. 2–3, pp. 73–272, 2015.

[2] M. Cavazza, F. Charles, and S. J. Mead, "Character-based interactive storytelling," *IEEE Intelligent Systems*, vol. 17, no. 4, pp. 17–24, 2002.

[3] A. Abdelreheem *et al.*, "PlaceIt3D: Language-guided object placement in real 3D scenes," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2025.

[4] B. M. Öcal, M. Tatarchenko, S. Karaoglu, and T. Gevers, "SceneTeller: Language-to-3D scene generation," in *Proc. European Conference on Computer Vision (ECCV)*, 2024, pp. 362–378.

[5] T. Brown *et al.*, "Language models are few-shot learners," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2020, pp. 1877–1901.

[6] Y. Zhou, I. Nuriddinov, A. E. Rhesa, W.-T. Lo, and T.-Y. Li, "RL-LABEL: Reinforcement learning-based label placement for dynamic AR scenes," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 6, pp. 2540–2552, 2023.

[7] M. Fiala, "ARTag: A fiducial marker system using digital techniques," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 590–596.

[8] C. Lee and A. Varshney, "Human-centric label placement in augmented reality," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 11, pp. 3660–3675, 2022.

[9] D. Xu, Y. Zhu, C. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5410–5419.

[10] U. Parihar *et al.*, "MonoPlace3D: Learning 3d-aware object placement for 3d monocular detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[11] W. Feng *et al.*, "LayoutGPT: Compositional visual planning and generation with large language models," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[12] C. Wang *et al.*, "Chat2Layout: Interactive 3D furniture layout with multimodal large language models," *arXiv preprint arXiv:2407.21333*, 2024.

[13] S. Peng *et al.*, "OpenScene: 3D scene understanding with open vocabularies," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 815–824.

[14] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "Affective computing and sentiment analysis," *IEEE Intelligent Systems*, vol. 31, no. 2, pp. 102–107, 2016.

[15] C. Chen *et al.*, "UniEmoX: Cross-modal semantic-guided large-scale pretraining for universal scene emotion perception," *IEEE Transactions on Image Processing*, 2025.

[16] G. Team and G. DeepMind, "Gemini: A family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[17] M. Kuyucu *et al.*, "Emotion recognition in virtual reality using sensor fusion with eye tracking," *Computers in Biology and Medicine*, vol. 197, p. 111070, 2025.