International Journal of Innovative Research in Engineering and Management (IJIREM)
ISSN (Online): 2350-0557, Volume 13, Issue 3, June 2026
DOI: https://doi.org/10.55524/ijirem.2026.13.3.15
Article ID IRP-1762, Pages 125-137
www.ijirem.org

SG-WORLD V2: A Planning-First Framework for Multimodal World Simulation

Piyush Thapliyal[1], Purva Mundada[2], Mukthikka V[3], and *Gurpreet Singh[4]

[1] BA Programme (NEP), School of Open Learning (SOL), University of Delhi, New Delhi, India
[2] MBA Scholar, Department of Management, JSPM University, Pune, India
[3] School of Aeronautical Engineering; Bharath Institute of Higher Education and Technology, India
[4] Endicott College of International Studies, Woosong University, Daejeon, South Korea

*Correspondence should be addressed to Gurpreet Singh ; gurpreetsinghmcse@gmail.com

Received: 20 May 2026; Revised: 7 June 2026; Accepted: 22 June 2026

ABSTRACT - Current systems that generate content across multiple modes, like video and audio, often face issues such as missing objects, weak connections between meanings, and poor coordination between visual and sound elements. These problems often arise because visual and audio parts are created separately, leading to mismatches between what is shown and what is heard. To tackle these issues, this paper introduces SG-WORLD V2, a framework that focuses on planning first for semantic reasoning and coordination in structured world simulations. Instead of directly creating video or audio, SG-WORLD V2 builds a semantic blueprint first. It translates natural language into machine-readable forms using reasoning based on the Universal Scene Graph (USG), adds a Quantification Rule to keep track of requested items; and uses Deterministic Pre-Temporal Synchronization (DPTS) to align audio and visual events during planning. The system also includes a self-correcting mechanism to check for semantic consistency before generating outputs. The framework was tested with 54 different prompts that cover object arrangements, spatial connections, actions, and synchronized audio-video situations. The results showed strong performance in keeping track of objects, maintaining relationships, and creating logically aligned plans. Overall, SG-WORLD V2 offers a clear and dependable semantic planning method that can support future systems for generating video and audio while improving consistency and alignment between different modes.

KEYWORDS— Audio-Visual Synchronization, Multimodal Generation, Semantic Planning, Universal Scene Graph.

I. INTRODUCTION

Generative Artificial Intelligence is progressing quickly, from producing content in a single format like images or text to creating more complex multimodal environments that combine different forms of media [1]. With the development of Diffusion Transformers (DiTs) and advanced Large Multimodal Models (LMMs), foundational architectures have been able to create visually appealing video sequences and contextually relevant audio from natural language descriptions. However, as multimodal generation becomes more standard, there are still significant technical challenges that need to be addressed. In real-world situations, visual and auditory information are closely linked, requiring sound to match spatial layout and actions to align with their corresponding sounds [1].

Despite the impressive quality of modern generative outputs, current systems still face issues such as semantic hallucinations, temporal inconsistencies, and major problems in cross-modal synchronization [1], [2]. The main limitation of existing multimodal systems is their fragmented pipeline structure. In many current frameworks, video and audio are treated as separate tasks, with the soundscape often generated as a post-production step, not naturally connected with visual elements [1]. This split can lead to "modality gaps", where the audio does not temporally or semantically match what is visible in the video. For example, systems might produce sounds like footsteps when no movement is visible or keep a sound effect even after the object causing it has left the scene [2]. Many existing systems use open-loop generation, creating media in one pass without strong mechanisms to check for semantic accuracy.

As a result, these systems often lack reliable methods for identifying and fixing semantic issues during generation. Another issue is compositional hallucination and object omission, which are ongoing problems in current diffusion-based systems. When dealing with complex prompts involving multiple entities, such as "a black dog chasing a rabbit", probabilistic models can miss key objects or misinterpret quantities. These issues are linked to weak reasoning abilities and poor grounding between entities [3]. While scene graphs, which represent objects and their relationships, have been used for visual understanding, their potential as a central framework for synchronized multimodal production is still not fully explored [4].

One key limitation found in this study is that current systems lack a unified semantic orchestrasssstion layer that can ensure deterministic synchronization and closed-loop multimodal reasoning. Models continue to struggle with capturing implicit relationships, such as hidden environmental interactions and contextual sounds needed for realistic simulations. Although recent innovations like FlowZero have introduced Dynamic Scene Syntax (DSS) to

improve spatial-temporal coherence using LLM-driven layouts [5] and systems like Anim-Director use LMM-powered agents for controlled script-to-video generation [6], few frameworks explicitly address deterministic synchronization between audio and video during the semantic planning stage.

To overcome these challenges, this paper presents SG-WORLD V2, a framework for controllable world simulation that focuses on multimodal reasoning and planning. Unlike traditional text-to-video generators that directly map language into pixels and sound waves, SG-WORLD V2 acts as a semantic reasoning core and autonomous planning layer. It breaks down natural language prompts into a Universal Scene Graph (USG) representation, forming a machine-readable blueprint for the generation process [4]. Operating as a high-level reasoning framework, SG-WORLD V2 provides semantic plans in structured JSON format and synchronization constraints that can guide various rendering tools like Sora, Stable Video Diffusion (SVD), or ElevenLabs. A major contribution of SG-WORLD V2 is Deterministic Pre-Temporal Synchronization (DPTS).

Traditional multimodal systems usually attempt synchronization during or after generation, which can lead to temporal drift [2]. In contrast, SG-WORLD V2 performs synchronization during the planning phase before rendering starts. Through DPTS, the framework creates an explicit synchronization events map, treating audio-visual timing as a logical constraint, not just a rendering artifact. This planning process supplies rendering systems with clear timing data, helping maintain better alignment between visual and audio elements.

The framework also introduces an explicit quantification mechanism to handle inconsistencies in object counts and compositional ambiguity. Unlike conventional black-box systems that might hallucinate or miss important elements, SG-WORLD V2 enforces clear Quantification Rules within its DSS planning framework. All visual entities are precisely defined, such as "one single dog" or "two distinct children," ensuring that every object from the scene graph is grounded in the semantic plan. This process improves compositional accuracy compared to standard prompt-to-media generation methods [3].

Another important part of the framework is Closed-Loop Pre-Production Verification [5]. SG-WORLD V2 includes a self-correcting critic that functions as a semantic quality assurance layer [6]. This module evaluates generated plans across five dimensions: object consistency, action consistency, spatial consistency, temporal consistency, and audio-visual consistency [1], [7]. By comparing the generated plans against the original prompts, the framework can detect omissions or logical errors and trigger an automatic re-review process before resources are used for rendering [5]. The effectiveness of SG-WORLD V2 was tested across 54 diverse prompts, including standard tests and user-defined stories. Experimental results showed consistent semantic planning behaviour, with the self-correction module successfully validating object relationships, synchronization, and cross-modal consistency across all scenarios. These findings provide evidence that integrating multimodal generation with structured scene-graph reasoning offers a reliable path toward realistic world simulation.

In summary, this research contributes the following to the field:

- **Deterministic Pre-Temporal Synchronization (DPTS):** A synchronization method that treats audio-visual timing as a logical constraint during planning to reduce temporal drift before generation.
- **Compositional Safeguarding through Explicit Quantification:** A semantic control mechanism that enforces clear numbering and grounding for all scene entities to prevent object omissions and hallucinations.
- **Closed-Loop Pre-Production Verification**: A self-correcting feedback system that checks multimodal plans against multiple semantic consistency criteria to enhance reliability.
- **Generator-Agnostic Multimodal Orchestration**: A universal semantic reasoning framework designed to control diverse generative models through structured, machine-readable JSON plans.

II. LITERATURE REVIEW

A. Introduction to Multimodal Generation and Its Evolution

Generative artificial intelligence has evolved from creating content in just one form, like text to image, to producing content that combines multiple forms, such as visual and audio information [5], [8]. Early breakthroughs in text-to-image (T2I) synthesis allowed the generation of images from natural language, but human experience is naturally multimodal. In real life, visual and auditory information are closely linked, such as the sound of footsteps matching a person's gait or dialogue matching lip movement [4], [9]. Because of this, multimodal generation is becoming more important in both research and industry. Video and audio are now produced together rather than in separate steps to ensure that the content aligns in meaning and feels immersive [2]. This shift is driven by the understanding that visual storytelling is closely connected to its auditory counterpart, and that systems must be able to process both forms of information simultaneously [2], [4].

B. Existing Text-to-Video Generation Methods

Text-to-video (T2V) generation has progressed from early methods that used Conditional Variational Autoencoders (CVAE) and Generative Adversarial Networks (GANs) to extract key visual features to more advanced models such as Diffusion Transformers (DiT) and space-time U-Net architectures. One development in this area is Sora, which uses a diffusion transformer to generate high-quality videos from text prompts by working with compressed latent representations [1], [7]. Models like Lumiere improve temporal smoothness without needing keyframes by using overlapping frame generation through a Space-Time U-Net (STUNet) architecture [7]. Although these models can produce visually impressive results, they are often computationally intensive and limited to short clips, making it difficult for them to maintain a consistent storyline across multiple scenes [6], [7]. Similarly, SEINE improves temporal continuity by employing a short-to-long diffusion strategy that enables smoother transitions between scenes in generated videos [9]. Zero-shot frameworks like FlowZero try to address these issues by using Large Language Models (LLMs) to generate structured Dynamic Scene Syntax

(DSS), which provides guidance for each frame to keep the video coherent over time [5].

C. Existing Text-to-Audio Generation Methods

Recent research in multimodal AI has focused on generating audio from text (TTA) and visual content (V2A). Models such as AudioLDM use latent diffusion techniques to turn text descriptions into audio waveforms. However, matching the generated audio to visual events is still a major challenge in multimodal systems [1], [2]. Models like T2AV integrate Audio-Visual ControlNets to combine visual and text information, making sure the sounds match what is seen [2]. Frameworks like Audio2AB demonstrate the value of trimodal generation by aligning facial expressions, gestures, and speech audio, showing the importance of "acoustic common sense" in creating realistic digital humans [10]. However, many of these systems are "reactive", trying to align audio after the video is created, which can lead to "sync drift" due to small delays during continuous generation [1], [2].

D. Scene Graph and Structured Semantic Representation Approaches

The research community has increasingly focused on structured representations, especially scene graphs, to address the limitations of models that treat elements as individual parts rather than connected components. For example, a model may struggle to distinguish between "an astronaut riding a horse" and "a horse riding an astronaut" [11]. A scene graph (SG) is a structured way to represent objects (nodes), their properties (attributes), and relationships (edges) in a scene, helping computers better understand and simulate the world [3], [7]. Tools like "Generate Any Scene" automatically create scene graphs to generate diverse and complex training data, showing that more detailed compositions improve the quality of generated content and prevent models from missing important elements from a prompt [3]. To connect different types of information across modalities, USG has been proposed as a way to align semantic meanings between text, images, video, and 3D data, offering a more unified approach for understanding complex scenes [4].

E. Cross-Modal Synchronization Challenges

Creating effective multimodal content requires close coordination between visual and audio elements. However, existing research points to a significant "modality gap," where audio does not match visual triggers. Humans are naturally sensitive to inconsistencies between what they see and hear, making precise synchronization one of the most important challenges in multimodal systems [1], [2]. Benchmarks such as T2AV-BENCH have introduced metrics like Frechet Audio-Visual Distance (FAVD) to measure how well these systems match the timing and content of audio and visual elements [2]. Synchronization becomes even more complex in longer content, where small delays can build up over time, leading to noticeable "sync drift" across extended sequences [1], [2]. Additionally, the lack of "acoustic common sense"—the intuitive understanding of how space and environment affect sound—can result in audio that feels disconnected from the visual context [1].

F. Limitations of Existing Systems

An analysis of current literature reveals several major flaws in existing frameworks. Most systems treat video and audio as separate tasks, creating noticeable "modality gaps" where sound is added after the video is completed [1], [7]. Many generative models also suffer from "compositional omission" and hallucinations because they lack strict rules for interpreting semantic instructions during the planning phase [3], [7]. Furthermore, many systems operate in an "open-loop" manner, generating media in one step without mechanisms to check whether the output matches the user's original intention [5], [6]. While some models, like Anim-Director, include self-reflection, they often rely on selecting the "best" result from several possible outputs, which is costly and lacks a clear logical structure [6].

G. Research Gap and Motivation for SG-World V2

The main research gap identified is that current multimodal systems do not have a unified semantic orchestration layer that can ensure deterministic synchronization and compositional verification before rendering starts [1]. Although FlowZero introduces Dynamic Scene Syntax for managing visual layouts [5], and Anim-Director uses agentic scripts for animation [6], neither of these frameworks treats audio-visual timing as a logical constraint that is defined during the planning stage. SG-WORLD V2 was created to overcome these shortcomings by introducing a structured planning layer that organizes scene information before the generation process begins. By using a Universal Scene Graph (USG) as the semantic foundation and implementing Deterministic Pre-Temporal Synchronization (DPTS), SG-WORLD V2 tackles these issues, thereby reducing hallucinations and synchronization errors by aligning visual and audio planning within the same structured representation [4].

H. Comparative Analysis of Existing Methods

Comparing existing multimodal generation and planning frameworks helps in understanding the strengths and weaknesses of current approaches. While recent models have made significant advances in generating images, videos, and audio, they still face several challenges, such as missing objects, incorrect relationships between objects, and poor semantic consistency [3]. In some cases, models struggle to fully comprehend how different objects relate to one another within a scene, which can negatively impact the quality of the generated output [11]. Another widespread issue is maintaining proper synchronization between visual and audio elements, which can result in mismatches between what is seen and what is heard [2]. Recognizing these limitations helps identify existing research gaps and emphasizes the need for better multimodal planning frameworks.

Table 1: Comparative analysis of representative multimodal generation and planning framework discussed in the literature. The comparison highlights their primary contribution, strengths, limitations, and relevance to the development of the proposed SG-WORLD V2 framework.

Model / Method	Modality	Main Contribution	Strengths	Limitations	Relevance to SG-WORLD
FlowZero [5]	Text-to-Video	Dynamic Scene Syntax (DSS) for spatio-temporal reasoning.	Improves motion smoothness and temporal coherence in zero-shot.	Focused on visual-only layouts; lacks integrated audio sync.	Motivates the use of DSS for structured planning.
Anim-Director [6]	Multimodal	Autonomous agent using LMMs for storyline-to-script workflow.	High versatility; maintenance of character consistency.	Relies on selecting the best from stochastic renders; high compute cost.	Validates the “Director Brain” concept for narrative management.
T2AV [2]	Text-to-Audio	Audio-Visual ControlNet for video-aligned sound generation.	Achieves state-of-the-art perceptual alignment.	Operates as an open-loop post-production aligner.	Identifies the need for synchronization benchmarks (T2AV-BENCH).
Generate Any Scene [3]	Visual	Scene graph driven data engine for systematic synthesis.	Ensures compositional coverage prevents object omission.	programmatically generated prompts can be unrealistic.	Establishes Scene Graphs as a requisite for reliable grounding.
Structure-CLIP [11]	Multimodal	Integration of Scene Graph Knowledge (SGK) into representations.	Solves "bag-of-words" failures in understanding relations.	Limited to evaluation/matching; not a generative planner.	Justifies structured grounding for fine-grained semantics.
Universal SG [4]	Multimodal	Unified taxonomy for semantics across text, video, and 3D.	Fully characterizes holistic scene semantics cross-modally.	Significant domain imbalance across datasets.	Informs the Universal Scene Graph core for cross-modal reasoning.
SEINE [9]	Text-to-Video	Short-to-long (S2L) diffusion model for transitions.	Enables smooth and creative scene-to-scene transitions.	Focused on autoregressive prediction rather than global logic.	Addresses the "shot level" limitation of current renderers.
Audio2AB [10]	Trimodal	Audio-driven collaborative generation of character animation.	Synchronizes face and gestures with speech audio.	Specialize in character animation only.	Highlights the importance of "acoustic common sense" in dynamics.

As shown in Table 1, current multimodal generation frameworks have made significant progress in specific tasks; however, their capabilities remain limited, as each method focuses on only one aspect of the multimodal generation and planning. While considerable progress has been made in areas like planning, scene comprehension, synchronization, and generation quality, most existing methods focus on just one part of the generation process [1], [7]. FlowZero enhances spatio-temporal planning, T2AV emphasizes audio-visual alignment, Structure-CLIP improves semantic understanding, and Universal Scene Graphs offer structured scene representations [2], [5], [11]. However, there is no single framework that integrates semantic grounding, object preservation, deterministic synchronization, and verification all within a unified process [1]. This gap drives the development of SG-WORLD V2, which brings together these features through a structured planning system based on a Universal Scene Graph [4].

I. Summary of Key Findings and Research Gap

The review of existing literature shows that multimodal generation has made significant strides in visual quality, scene understanding, audio creation, and planning-based control. Today’s diffusion and transformer-based models can produce highly realistic outputs, but several key challenges remain unsolved [1], [7]. Current systems still face issues such as missing objects, semantic inconsistencies, incorrect object relationships, and poor grounding of complex prompts [3], [11]. Moreover, many frameworks create visual and audio content separately, leading to synchronization issues and reduced perceptual realism [1], [2].

Recent planning-based and agent-driven methods aim to enhance control through scene layouts, iterative improvements, and reasoning-based workflows [5], [6]. However, these approaches usually concentrate on specific aspects of the generation process rather than offering a single, integrated solution for semantic planning, synchronization, verification, and object-level control [1]. Consequently, many systems still work in an open-loop fashion, where generated outputs are not systematically checked against the original prompt requirements [6].

From these observations, it is clear that there is a need for a generator-agnostic orchestration framework that can combine structured scene understanding, deterministic planning, temporal synchronization, and self-correction in one process [1], [6]. Such a framework would not only

enhance generation quality but also improve reliability, consistency, and control in multimodal tasks [3], [5]. To fill this gap, the proposed SG-WORLD V2 framework employs a Universal Scene Graph as its core semantic structure [4], allowing for object-level reasoning, precise quantification, synchronized event planning, and logic-based verification before any rendering occurs [3], [5]. This approach moves multimodal generation from a purely probabilistic method toward structured and verifiable world simulation [1].

III. PROPOSED METHODOLOGY

A. System Overview

The SG-WORLD V2 framework is built as a semantic planning system for simulating a multimodal world. It functions as a "director brain" that interprets the user's input and develops a structured plan before any video or audio content is created. Rather than directly producing video or sound, It first plans the scene [5], [6]. It takes the input prompt and identifies key components like objects, actions, relationships, and environmental details, then arranges them in a structured way [3], [4]. This ensures that both visual and audio elements remain properly aligned within the same scene plan [1], [4]. By following a planning-first method, the framework helps reduce common issues found in multimodal systems, such as missing items, mismatched scene details, or improper timing between visual and audio events [2], [3], [7]. Since the framework focuses on planning instead of direct generation, its output can be used with various video and audio creation systems [5], [6]. The full workflow of SG-WORLD V2 is shown in Figure 1.

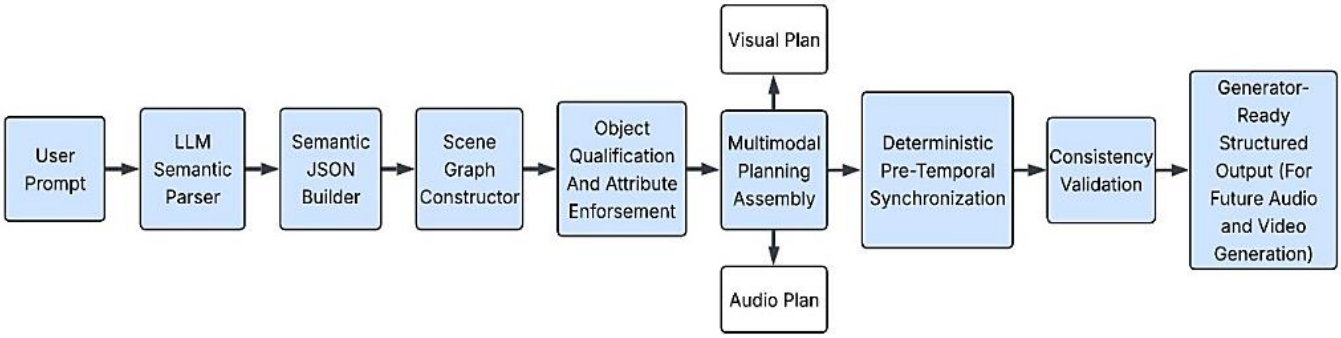

Figure 1: Overall architecture of the SG-WORLD V2 framework, showing complete semantic planning, synchronization, verification, and output pipeline.

B. Natural Language Semantic Parsing

The first step of the process uses a Large Multimodal Model (LMM) serving as a "semantic reasoning core" to extract basic elements from the input text [6], [12]. Traditional models often struggle with a "bag-of-words" issue, which makes them unable to recognize subtle differences in relationships [11]; SG-WORLD V2 addresses this by identifying specific entities, actions, descriptive features, and environmental factors [4], [11], [12]. The semantic parsing process is defined as:

$$, \quad P \rightarrow S = \{O, At, D, R\} \qquad (1)$$

In this expression, P represents the input user's prompt, which is converted into a structured semantic representation S containing detected objects O, attributes At, actions D, and relationships R. Breaking the prompt into these components helps the framework identify what is present in the scene and how different elements are connected. This organized structure acts as the basis for the planning steps that come next.

C. Scene Graph Construction

Based on the parsed semantic data, the framework transforms the extracted information into a structured Scene Graph. This graph acts as the semantic foundation of the scene by organizing objects and their relationships into a format that can be understood by machines [3], [4]. Formally, the scene is represented as:

$$G = (V, E) \qquad (2)$$

Where G stands for the scene graph, V is the set of object nodes, and E represents the connections between these objects. These connections can describe spatial layouts, interactions, or other meaningful relationships within the scene [3], [4]. By utilizing a USG, the framework ensures that both visual and auditory elements are based on the same interconnected structure. An example of a scene graph is shown in Figure 2 [4].

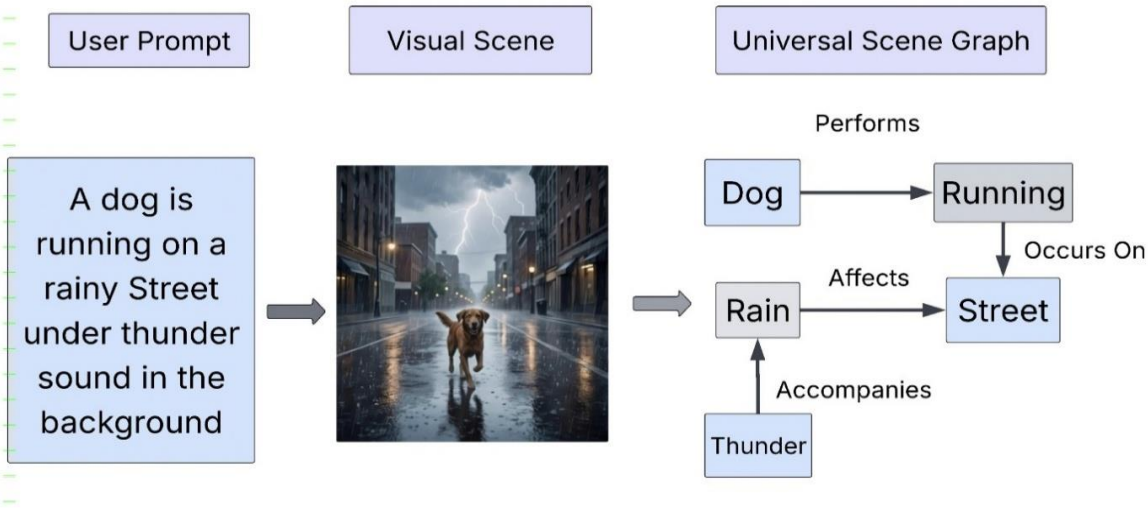

Figure 2: Example of Universal Scene Graph (USG) Construction from a Natural Language Prompt.

D. Object Qualification and Attribute Enforcement

A major technical challenge in present generative models is compositional omission, where requested objects are either left out or inaccurately depicted in complex scenes [3]. Additionally, modern diffusion-based systems often have difficulty maintaining numerical constraints and detailed object attributes when prompts involve multiple entities or intricate relationships [11]. To overcome these issues, SG-WORLD V2 includes a structured object qualification and quantification phase during semantic planning. Each object retrieved from the Scene Graph is given clear descriptive attributes and quantity-related constraints before multimodal planning starts. This process transforms natural language descriptions into structured semantic constraints, enhancing object preservation, semantic consistency, and subsequent verification.

$$Q(o_i) = \{a_1, a_2, \ldots, a_n\} \quad (3)$$

Here, Q represents the qualification function applied to each object node o_i in the Scene Graph. The function assigns a structured set of attributes $\{a_1, a_2, ..., a_n\}$, which may include features such as color, material, size, motion state, environmental conditions, and spatial location. In addition to assigning attributes, the framework also retains quantity-related information linked to each object throughout semantic planning. Together, these constraints aid in ensuring object accuracy and consistency prior to multimodal planning and synchronization.

E. Multimodal Planning Assembly

In this module, SG-WORLD V2 converts a static scene graph into dynamic, modality-specific planning instructions. The goal of this stage is to turn the structured semantic data into a clear multimodal plan that shows how the scene should be visually and acoustically presented. As a semantic planning and coordination layer, the framework keeps track of object relationships, contextual features, and time-based connections while changing natural language instructions into structured, machine-readable planning data. This allows the scene to be thoroughly examined, validated, and represented with high semantic accuracy before any further processing or execution.

- **Visual Planning**

The visual planning part outlines what needs to be shown visually in the scene. It covers object positions, movement paths, environmental settings, camera angles, scene layout, and visual actions, ensuring that all objects from the semantic data are fully included. Each visual element is directly linked from the scene graph to a structured plan, making sure that the final visual outcome matches the original user request in meaning.

- **Audio Planning**

The audio planning part determines what should be heard along with the visual scene. It identifies sound events, background noise, sound volume, and the timing of sounds. A “proximity rule" is used where close visual events are paired with strong audio cues, while faraway elements are presented as background sounds. This method ensures that the audio elements are properly connected to the visual parts of the scene.

For example, with the prompt “A dog is running on a rainy street,” SG-WORLD V2 creates a structured multimodal plan. The visual part includes the dog’s movement, street view, rain intensity, and the shiny wet road. The audio part includes the regular sound of footsteps in water and the ongoing background of rain. These structured plans create a single, unified multimodal blueprint that accurately and clearly represents the scene. Figure 3 shows an example of a multimodal plan created by the framework.

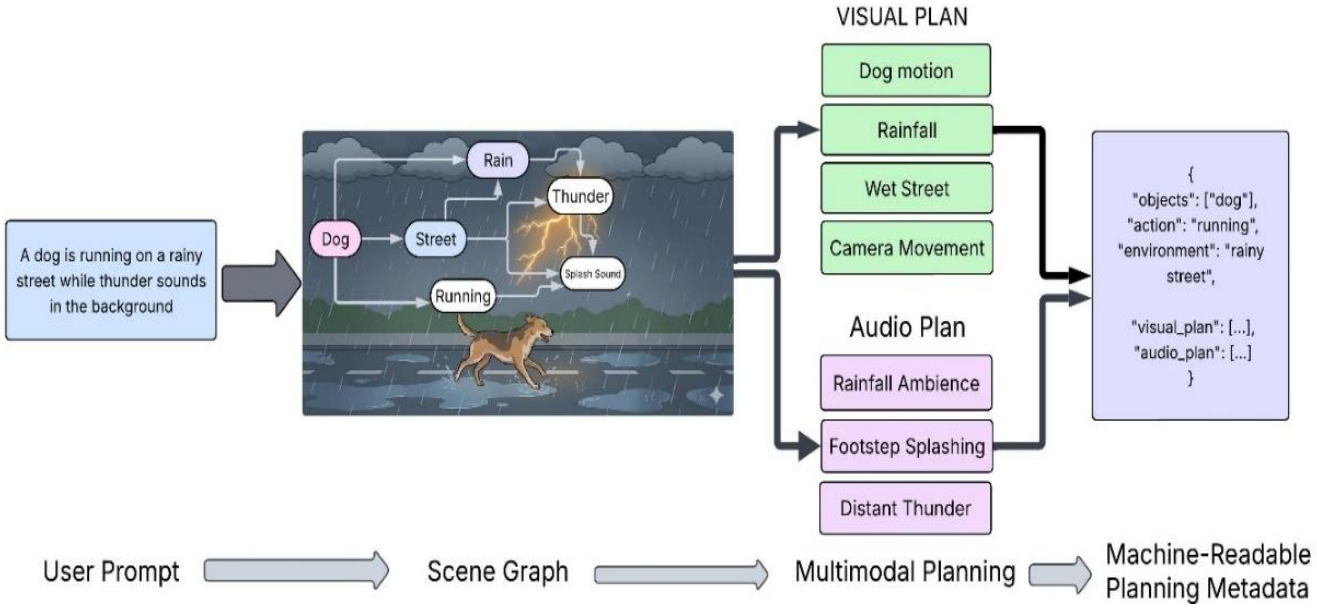

Figure 3: Example of SG-WORLD V2 multimodal planning process showing the transformation of a user prompt into a scene graph, visual plan, audio plan, and machine-readable planning metadata

F. Deterministic Pre-Temporal Synchronization (DPTS)

Unlike systems that adjust audio after the video is created, SG-WORLD V2 uses Deterministic Pre-Temporal Synchronization (DPTS). In DPTS, timing is considered as a planning constraint, where visual events are connected to their matching audio triggers before the generation process starts. The synchronization mapping is shown as:

$$T = \{(e_i, t_i) \mid i = 1, \ldots, n\} \quad (4)$$

Here, T is the temporal synchronization map. Each pair links a planned event e_i with its assigned time t_i, creating a shared timeline for both visual and audio elements [2]. This method ensures that visual and audio elements stay aligned before the final output is generated, which helps prevent synchronization issues often seen in separate multimodal systems. The process of aligning temporal events is shown in Figure 4.

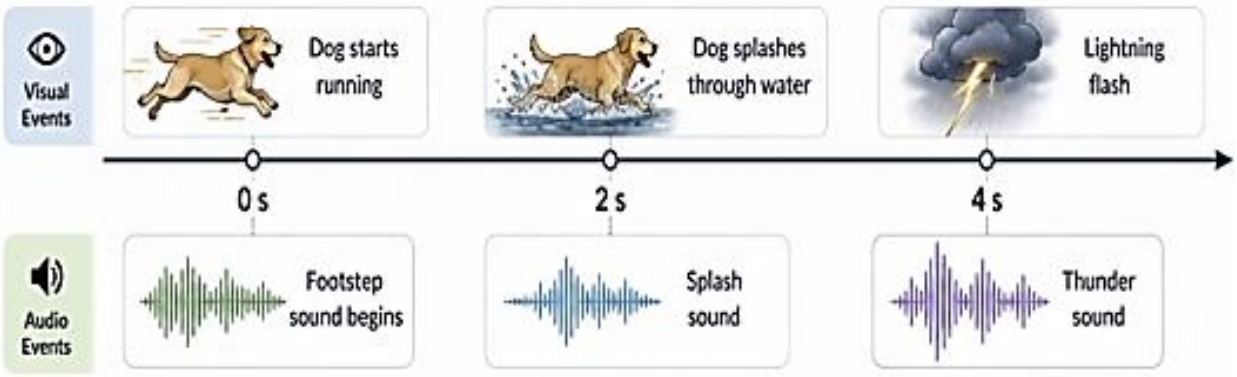

Figure 4: Deterministic Pre-Temporal Synchronization (DPTS) example showing visual events and corresponding audio events before generation

G. Closed-Loop Pre-Production Verification

The framework includes a self-correcting feedback system that checks the generated plans against the original scene graph based on five consistency factors: object consistency, action consistency, spatial alignment, temporal alignment, and audio-visual synchronization. The validation process is expressed through the formula:

$$C = \frac{M}{N} \quad (5)$$

In this formula, C represents the overall consistency score of the plan. M is the number of constraints that are met, while N is the total number of constraints taken from the original scene graph. The verification process reviews five main aspects of the plan: object consistency, action consistency, spatial consistency, temporal consistency, and audio-visual synchronization. A higher score shows that the generated plan is closer to the original user instruction. Unlike many existing multimodal systems that go directly from interpreting the prompt to generating content, SG-WORLD V2 includes a verification step before the generation begins. If the framework finds an object missing, an incorrect action, a spatial mismatch, a timing issue, or a synchronization problem, the plan is reviewed and modified

before moving to the next step. Because these corrections are made during the planning phase rather than after video or audio generation, they can be done more efficiently while keeping the original meaning of the prompt intact. Currently, all constraints are treated the same when calculating the consistency score. Future improvements may consider using weighted scoring, where certain important scene elements are given more weight than minor attribute differences.

H. Generator-Agnostic Output Layer

The last stage of the methodology creates a high-precision, machine-readable JSON file that includes the verified visual sequence, acoustic plan, and DPTS event map. As the framework is generator-agnostic, this structured script functions as a universal set of instructions that can be converted into specific API commands for any external rendering system, such as Sora, Lumiere, or Stable Video Diffusion [1]. This design separates the semantic planning from the rendering process, allowing the framework to be compatible with various generation engines without modifying its core logic. As a result, SG-WORLD V2 can act as a flexible base for future multimodal world-building systems.

IV. EXPERIMENTAL SETUP AND EVALUATION FRAMEWORK

A. Experimental Design

The experimental evaluation of SG-WORLD V2 was done to assess its effectiveness as a semantic planning and coordination system. Unlike traditional multimodal models that directly produce images, videos, or audio. It emphasizes planning before the generation phase. Therefore, the experiments mainly measured how well the framework translates natural language prompts into structured, synchronized, and logically consistent multimodal plans. The evaluation focused on five main areas: object composition, relational reasoning, attribute preservation, temporal synchronization, and overall multimodal consistency. These areas were chosen to determine whether the generated plans stay true to the original prompt while maintaining semantic consistency throughout the entire planning process.

B. Prompt Dataset and Scenario Categories

SG-WORLD V2 was tested using a dataset of 54 prompts, which included 4 foundational prompts used during initial testing and 50 user-defined multimodal prompts created to test more complex reasoning in different scene situations. To ensure a wide range of evaluation, the prompts were designed to test semantic, spatial, temporal, and multimodal reasoning across various scene types. The evaluation dataset is summarized in Table 2.

Table 2: Distribution of Prompt Categories Used in the SG-WORLD V2 Evaluation

Scenario Category	Description	Number of Prompts
Foundational Logic Validation	Initial prompts used to verify basic pipeline flow, JSON extraction, and core Universal Scene Graph (USG) construction [4].	4
Multiple Objects and Composition	Prompts designed to test the Quantification Rule by requiring accurate counting and preservation of multiple distinct entities within a scene [3], [5].	12
Complex Spatial Relations	Prompts involving subject–relation–object structures, object placement, and spatial dependencies between entities [4], [11].	10
Dynamic Actions and Environmental Conditions	Prompts containing motion, interaction, and environmental context such as rain, wind, or movement-based scene changes requiring semantic expansion [5].	15
Synchronized Audio-Visual Events	Multimodal prompts requiring structured temporal alignment between visual actions and acoustic events such as thunder, footsteps, splashing, or ambient sound.[2]	13
Total	Total prompts evaluated in the experimental framework	54

The prompts were organized into four primary scenario groups:

- **Multiple Objects and Composition:** This group checks if the framework can accurately represent several objects in the same scene, especially when the prompt specifies a certain number of items. This group helps in identifying issues related to missing objects or incorrect preservation of objects [3].
- **Complex Spatial Relations:** This group evaluates whether SG-WORLD V2 can correctly interpret the relationships and positioning of objects in a scene.For instance, a prompt like "A blue backpack resting on a wooden chair" requires that the relationship between the backpack and the chair is accurately maintained.
- **Dynamic Actions and Environmental Conditions**: This group includes prompts that involve movement, interactions, and environmental settings.
- An example is "A dog is running on a rainy street while thunder sounds in the background." These prompts test if the framework can coordinate actions with the environment.
- **Synchronized Audio-Visual Events**: This group evaluates if the framework can logically match sound events with visual actions using structured time planning via Deterministic Pre-Temporal Synchronization (DPTS).

This distribution makes sure that the evaluation includes a wide variety of situations involving semantic, spatial, temporal, and audio-visual reasoning.

C. Evaluation Metrics

To assess the performance of SG-WORLD V2, a set of validation metrics was developed to determine how well the generated plans maintain the meaning of the original prompts. The evaluation metrics and their purposes are summarized in Table 3. These selected metrics examine various aspects of semantic planning, such as object preservation, attribute accuracy, relational reasoning, temporal coordination, and overall consistency across different modes.

Table 3: Evaluation Metrics Used For SG-WORLD V2 Validation

Metric	Purpose	Evaluation Focus
Object Completeness	To verify that all entities mentioned in the input prompt are preserved throughout the planning pipeline.	Detection of omitted objects, missing entities, or incorrect object counts using the Quantification Rule.
Attribute Accuracy	To ensure that object-specific attributes are correctly assigned during semantic planning.	Validation of color, material, size, state, motion, and environmental properties within the generated specifications [3].
Relational Consistency	To verify that semantic relationships between objects remain preserved throughout scene graph construction and planning [4].	Maintenance of spatial, functional, and behavioural relationships defined in the Universal Scene Graph (USG) [4],[3], [7].
Temporal Alignment	To evaluate whether planned visual and audio events are correctly ordered over time [2].	Verification of chronological event sequencing and temporal synchronization across multimodal outputs.
Audio-Visual Semantic Coherence	To measure whether planned audio events logically match the visual context of the scene [1].	Validation of sound-event appropriateness for corresponding visual actions or environments, such as rainfall, footsteps, or ambient sound.
Overall Consistency Score (C)	To provide a unified quantitative measure of overall semantic reliability.	Computed as C = M/N, where M is the number of satisfied semantic constraints and N is the total number of constraints extracted from the Universal Scene Graph (USG) [4].

The evaluation metrics include:

- **Object Completeness:** This measures whether all objects mentioned in the prompt are included in the final planning output.
- **Attribute Accuracy:** This measures whether object properties such as color, size, material, motion, and environmental details are correctly assigned.
- **Relational Consistency:** This measures whether the relationships between objects—such as position or interaction—are maintained accurately throughout the planning process.
- **Temporal Alignment:** This measures whether visual and audio events are placed in the correct sequence and aligned properly over time.
- **Audio-Visual Semantic Coherence:** This measures whether the planned sounds match the visual scene.For example, rainfall should correspond to a rainy environment, and splash sounds should align with movement on a wet surface. [15]

Together, these metrics offer a complete evaluation of how well meaning is preserved and how consistent different types of information are throughout the planning process. The primary numerical measure used in the assessment is the Overall Consistency Score (C):

$$C = \frac{M}{N} \qquad (6)$$

Here, C represents the consistency score, M is the number of satisfied semantic constraints, and N is the total number of constraints extracted during the Universal Scene Graph (USG) construction phase [4]. This score reflects how well SG-WORLD V2 maintains the meaning of the original prompt throughout the planning process. A higher consistency score indicates a stronger match between the input prompt and the generated planning output. The framework includes the preservation of key semantic elements such as object presence, attributes, relationships between entities, and temporal alignment across planned multimodal events.

D. Qualitative Validation and Comparative Analysis

The generated plans were examined manually and compared with the original prompts to assess how well the meaning was preserved, the quality of the planning, and the overall consistency. This review checked whether SG-WORLD V2 maintained the intent of the prompt throughout the stages of semantic parsing, scene graph creation, multimodal planning, and timing coordination. The framework was also compared with traditional systems that rely solely on prompts to generate multimodal outputs. These systems often struggle with issues such as missing objects, wrong attributes, misunderstandings of meaning, or poor timing between sounds and actions [3]. The findings show that SG-WORLD V2 provides a more organized planning process by first gathering information about objects, their connections, characteristics, and timing before generating the final output. This helps ensure that the final planning stays closer to the original prompt. In particular, the framework's Quantification Rule helps keep track of important elements in complex scenarios, improving the accuracy of object representation.

E. Implementation Output Representation

For this evaluation, the outputs were analyzed as structured semantic data instead of as final media files such as images, videos, or audio. This matches the goal of SG-WORLD V2, which is to understand and organize scene details before

creating multimedia content. The main types of output include:

- **Universal Scene Graph (USG):** JSON-based structures made up of nodes and edges that describe scene meaning, object relationships, and overall layout [4]
- **Multimodal Specification Plans:** Organized plans that show object actions, scene details, environmental context, and descriptions of sounds.
- **Temporal Synchronization Maps:** Structures that link events to specific times to ensure timing consistency in planned interactions.
- **Self-Correction Metadata:** Logs generated during internal consistency checks across all 54 prompts. These logs note whether each plan passed or failed checks related to object completeness, meaning alignment, attribute accuracy, and timing consistency.

Although the evaluation showed promising results, there are still some limitations. The current study only assessed the plans created by SG-WORLD V2 and not the final videos or audio generated by other systems. Therefore, the reported consistency scores reflect the quality of the plans themselves, not the final multimedia content. Future work will involve testing the framework with different video and audio generation models to see how effectively the planning outputs can be turned into complete multimedia experiences. [17]

V. RESULTS

A. Overall Performance Analysis

The experimental evaluation of SG-WORLD V2 was conducted using a dataset of 54 prompts, which included 4 foundational test prompts and 50 user-defined multimodal scenarios. The results show that the framework functioned well as a semantic planning and orchestration system across various types of prompts. All the prompts evaluated were successfully processed through the entire pipeline. None of the prompts encountered failures during semantic parsing, scene graph construction, multimodal planning, or consistency verification.[13]

It also demonstrates that this model is capable of transforming natural language prompts into structured multimodal plans while preserving logical consistency throughout the planning process. Because the framework separates semantic reasoning from direct media generation, it addresses many of the common problems found in end-to-end multimodal generation systems, such as missing objects, incorrect relationships, or mismatched outputs.

Across the evaluation set, the framework produced structured JSON outputs and planning metadata for all tested prompts. The generated plans were logically organized throughout the pipeline, from semantic parsing to multimodal planning and validation. These findings suggest that SG-WORLD V2 can act as a dependable planning layer before video or audio generation, making it a good candidate for integration with future multimodal generation systems. The quantitative performance of SG-WORLD V2 across the assessed semantic planning dimensions is summarized in Table 4. The results indicate strong performance in object preservation, attribute mapping, relational reasoning, temporal alignment, and audio-visual semantic coherence across the 54 evaluated prompts.

Table 4: Quantitative Performance Metrics of SG-WORLD V2 (N = 54)

Metric	Successful Cases	Percentage (%)	Derivation Method
Object Completeness	54	100.0%	Verification of object preservation and quantification throughout the planning pipeline.
Attribute Accuracy	52	96.3%	Validation of semantic attribute mapping in generated multimodal plans.
Relational Consistency	51	94.4%	Internal verification of spatial, functional, and behavioural relationships within the Universal Scene Graph (USG)
Temporal Alignment	52	96.3%	Analysis of event-to-time mappings generated by the DPTS module.
Audio-Visual Semantic Coherence	51	94.4%	Verification of contextual alignment between planned sound events and visual scene descriptions.
Overall Consistency Score	52	96.3%	Average consistency achieved across all semantic validation dimensions.

B. Object Preservation and Attribute Accuracy

A key objective of SG-WORLD V2 is to preserve the original meaning of the user's prompt as faithfully as possible. [19] This was evaluated using Object Completeness and Attribute Accuracy. The results indicated that objects identified during semantic parsing were consistently reflected in the final planning output. For instance, prompts with a single object, multiple objects, or scenes with detailed descriptions were accurately represented in the generated plans without significant loss of information.

The framework also preserved object-specific attributes with high consistency. Characteristics such as color, size, material, movement, and environmental conditions remained consistent with the original prompt across all tested scenarios. Examples include a red car moving through a street scene, rainfall on a road, and environmental settings like an old library or a busy city street. [14]

These results indicate that the Quantification Rule and attribute qualification stages help prevent the omission of objects and enhance semantic consistency, especially in prompts involving multiple entities or complex scene descriptions. To assess performance across various types of

prompts, the test dataset was categorized into different scenario groups. Table 5 provides a summary of the consistency scores and success rates for each category.

Table 5: Performance by Scenario Category

Scenario Category	No. of Prompts	Consistency Score (%)	Success Rate
Foundational Logic Validation	4	100.0%	4 / 4
Multiple Objects and Composition	12	98.5%	12 / 12
Complex Spatial Relations	10	94.8%	10 / 10
Dynamic Actions and Environmental Conditions	15	95.7%	15 / 15
Synchronized Audio-Visual Events	13	94.2%	13 / 13
Total	54	96.3%	54 / 54

C. *Relational and Spatio-Temporal Consistency*

SG-WORLD V2 uses the Universal Scene Graph (USG) as the main structure for organizing objects, actions, and relationships. Based on the evaluation, this design helped the framework maintain consistency in scene structure across various types of prompts [4].

The framework effectively managed complex spatial relationships, including how objects are placed, how entities interact with each other, and how actions move through a scene. For instance, prompts that describe one object resting on another, movement through an environment, or multiple interacting entities were converted into structured plans while keeping their intended relationships intact. This was particularly useful in dynamic scenes that involve movement over time. Actions like walking, running, flying, or interacting with surrounding objects remained logically connected throughout the planning sequence. These findings suggest that using a scene graph for planning offers a structured way to retain both spatial and temporal information from the original prompt, helping maintain consistency throughout the planning process.

D. *Cross-Modal Synchronization and Audio-Visual Coherence*

SG-WORLD V2 is designed to plan both visual and audio information together before the generation process starts. This is done through Deterministic Pre-Temporal Synchronization (DPTS), which connects sound events with visual actions during the planning stage, rather than adding audio after video generation. During evaluation, the framework maintained a logical alignment between planned visual events and their corresponding audio descriptions across different types of prompts. For example, in prompts describing a dog running on a rainy street, the framework planned visual movement along with rainfall ambience, footstep splashes, and background thunder.[16]

This resulted in a coherent multimodal plan in which the audio naturally matched the visual scene. These results suggest that planning synchronization before generation can offer a more reliable and interpretable method for maintaining alignment between visual and audio elements [1]. Although the evaluation was based on planning outputs rather than the final generated media, the findings indicate that early synchronization planning can help maintain stronger consistency between visual and audio components.

VI. DISCUSSION

A. *Comparison with Existing Work*

SG-WORLD V2 offers a different method compared to many existing multimodal generation systems, such as Sora, Lumiere, and AudioLDM [1]. Most of these systems work by directly converting text into video or audio. Although these models can create impressive content, previous studies have found that they may still have problems like missing objects, incorrect relationships between objects, and poor synchronization between sound and motion [3], [11].

SG-WORLD V2 takes a different approach. Instead of immediately generating video or audio after receiving a prompt, it first creates a structured semantic plan using a Universal Scene Graph (USG) [4]. This planning-first method helps organize objects, actions, relationships, and timing before the actual generation begins. As a result, common issues like missing objects, incorrect relationships, and synchronization problems can be detected and addressed earlier in the process, rather than after the content has been produced. [18]

B. *Interpretation of Results and Key Strengths*

The evaluation across 54 prompts showed that SG-WORLD V2 consistently maintained scene structure, object relationships, and multimodal alignment. The findings suggest that the framework works well as a semantic planning system for building multimodal worlds. One major strength of the framework is the use of USG reasoning, which gives a shared semantic structure for representing scene elements and their relationships [4]. This helps ensure consistency between visual and audio planning outputs.

The Quantification Rule also plays an important role by keeping track of the number of objects mentioned in the prompt, reducing the chance of missing entities in complex scenes. Additionally, Deterministic Pre-Temporal Synchronization (DPTS) improves alignment between visual and audio planning by assigning timing relationships during the planning phase itself.

Finally, the self-correction verification loop enhances reliability by checking the generated plan against the original prompt and identifying inconsistencies before the final output is created. Together, these elements demonstrate that a planning-first framework can improve control, interpretability, and semantic consistency in multimodal AI systems.

Overall, the results suggest that structured planning can enhance semantic consistency in multimodal systems. Across various types of prompts, SG-WORLD V2 was able to retain important scene information while maintaining alignment between visual and audio planning components. These findings support the idea that planning before generation can offer better control and interpretability than relying solely on direct prompt-to-generation methods.

C. *Theoretical Implications*

From a theoretical standpoint, SG-WORLD V2 introduces a planning-first framework for multimodal AI that views

scene generation as a structured reasoning task, rather than just a generation task. The framework suggests that challenges like semantic inconsistency, object omission, and audio-visual misalignment can be more effectively addressed through explicit planning before media generation starts. By using a Universal Scene Graph (USG) as a common semantic representation, SG-WORLD V2 links visual and audio elements through the same structural logic, instead of treating them as separate outputs [4].

This work contributes to multimodal research by suggesting that semantic planning should be viewed as a separate layer between user prompts and media generation. Rather than relying entirely on the generation model to understand and organize a scene, the planning layer offers a structured representation that can improve control, transparency, and consistency in multimodal applications. [20]

D. Limitations of the Current Framework

Despite these advantages, SG-WORLD V2 has several current limitations. At the moment, the framework focuses on semantic planning, orchestration, and consistency verification. It creates structured JSON planning metadata, but it does not directly produce video frames or audio outputs. The experimental evaluation was done using a dataset of 54 prompts. While this was helpful for testing scene reasoning, object relationships, and synchronization logic, broader testing with larger and more diverse sets of prompts would provide stronger evidence of the framework's performance in real-world situations.

Another limitation is its reliance on large language model reasoning, which may bring practical challenges like longer inference times, higher computational costs, and dependency on the performance of external models. Additionally, the current evaluation focuses on the quality of the generated plans rather than the quality of the final rendered media. Although the framework produces structured and logically consistent planning outputs, further experiments are needed to assess how well these plans translate into real video and audio generation systems.

One more limitation is that the current evaluation was done on planning outputs rather than human-rated multimedia outputs, which makes direct comparison with end-to-end generation systems difficult. Future validation using multiple rendering backbones would offer stronger evidence of the framework's practical effectiveness. [17]

E. Future Research Directions

Future work will aim to expand SG-WORLD V2 into broader multimodal workflows. One key direction is connecting the planning framework with downstream video and audio generation systems through a modular interface, allowing structured plans to be automatically converted into rendered outputs. Another area of focus is extending the framework from single-scene planning to multi-scene storytelling, where object continuity, character consistency, and environmental transitions can be maintained across longer sequences [12].

Further improvements may include richer audio-environment modeling, such as simulating how sound behaves based on factors like distance, material, and space. In addition, larger-scale testing and evaluations based on human feedback can help improve the reliability and practical usability of the framework in future creative and research applications. Future studies may also explore weighted verification strategies, where critical elements like primary subjects, key actions, and synchronization events are given higher priority during validation. This could further enhance the robustness of the self-correction process and provide a more detailed assessment of semantic consistency in complex multimodal scenarios. [21]

VII. CONCLUSION

This study tackled the issues of semantic inconsistency and audio-visual misalignment in multimodal AI systems [3]. To address these challenges, SG-WORLD V2 was introduced as a planning-first semantic orchestration framework that organizes multimodal content prior to the generation process. By integrating Universal Scene Graph (USG) reasoning, the Quantification Rule, Deterministic Pre-Temporal Synchronization (DPTS), and structured JSON-based planning, the framework offers a unified way to represent visual and audio elements [4]. Testing across 54 prompts showed that SG-WORLD V2 successfully maintained object details, preserved semantic connections, and enhanced cross-modal alignment during the planning phase. The results indicate that structured planning can enhance consistency, control, and transparency in multimodal systems. Future efforts will involve larger-scale testing and integration with video and audio generation models to enable fully automated multimodal content production. By incorporating structured semantic planning before content generation, SG-WORLD V2 illustrates how multimodal systems can gain from explicit reasoning and verification prior to content creation, resulting in more controllable, understandable, and dependable world simulations. [22]

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

ACKNOWLEDGMENT

The authors would like to express sincere gratitude to the supervisor and faculty members for their guidance, encouragement, and constructive feedback throughout this research work.

REFERENCES

[1] Z. Li et al., "Multimodal Video Generation Models with Audio: Present and Future," 2026. Available from: http://doi.org/10.13140/RG.2.2.26531.31528.

[2] S. Mo, J. Shi, and Y. Tian, "Text-to-Audio Generation Synchronized with Videos," arXiv:2403.07938, Mar. 2024. Available from: http://doi.org/10.48550/arXiv.2403.07938.

[3] Z. Gao, W. Huang, J. Zhang, A. Kembhavi, and R. Krishna, "Generate Any Scene: Scene Graph Driven Data Synthesis for Visual Generation Training," arXiv:2412.08221, Mar. 2026. Available from: http://doi.org/10.48550/arXiv.2412.08221.

[4] S. Wu, H. Fei, and T.-S. Chua, "Universal Scene Graph Generation," arXiv:2503.15005, Mar. 2025. Available from: http://doi.org/10.48550/arXiv.2503.15005.

[5] Y. Lu, L. Zhu, H. Fan, and Y. Yang, "FlowZero: Zero-Shot Text-to-Video Synthesis with LLM-Driven Dynamic Scene Syntax," arXiv:2311.15813, Nov. 2023. Available from: http://doi.org/10.48550/arXiv.2311.15813.

[6] Y. Li et al., "Anim-Director: A Large Multimodal Model Powered Agent for Controllable Animation Video

Generation," arXiv:2408.09787, Aug. 2024. Available from: http://doi.org/10.48550/arXiv.2408.09787.
[7] A. Fime, S. Mahmud, A. Das, M. S. Islam, and H.-H. Kim, "Automatic Scene Generation: State-of-the-Art Techniques, Models, Datasets, Challenges, and Future Prospects," arXiv:2410.01816, Sep. 2024. Available from: http://doi.org/10.48550/arXiv.2410.01816.
[8] Y. Li, M. R. Min, D. Shen, D. Carlson, and L. Carin, "Video Generation From Text," arXiv:1710.00421, Oct. 2017. Available from: http://doi.org/10.48550/arXiv.1710.00421.
[9] X. Chen et al., "SEINE: Short-to-Long Video Diffusion Model for Generative Transition and Prediction," arXiv:2310.20700, Nov. 2023. Available from: http://doi.org/10.48550/arXiv.2310.20700.
[10] L. Niu, W. Xie, D. Wang, Z. Cao, and X. Liu, "Audio2AB: Audio-driven collaborative generation of virtual character animation," *Virtual Reality & Intelligent Hardware*, vol. 6, no. 1, pp. 56–70, Feb. 2024. Available from: http://doi.org/10.1016/j.vrih.2023.08.006.
[11] Y. Huang et al., "Structure-CLIP: Towards Scene Graph Knowledge to Enhance Multi-modal Structured Representations," arXiv:2305.06152, Dec. 2023. Available from: http://doi.org/10.48550/arXiv.2305.06152.
[12] Tian, "A Large Language Model-Based System for Semantic Understanding and Automated Scene Generation in Animation Scripts," in *Proceedings of the 2nd International Conference on Machine Intelligence and Digital Applications (MIDA '25)*, Jul. 2025, pp. 116–120. Available from: https://doi.org/10.1145/3744464.3744483.
[13] "Multimodal and Large Language Model Approaches in Cybersecurity: A Systematic Review," *International Journal of Engineering, Science and Environment (IJESE)*, vol. 1, no. 1, 2026. Available from: https://ijese.in/journal/article/view/18.
[14] Singh, A. Naaz, A. Syed, and V. Akhila, "AI-Assisted Storytelling: Enhancing Narrative Creation in Digital Media," *Preprints*, 2026. Available from: https://doi.org/10.20944/preprints202601.0330.v1.
[15] Singh, L. Qamar, N. V. Volta, A. Velamuri, and A. Khanyile, "Vision–Language Foundation Models and Multimodal Large Language Models: A Comprehensive Survey of Architectures, Benchmarks, and Open Challenges," *Preprints*, 2026. Available from: https://doi.org/10.20944/preprints202602.0467.v2.
[16] Singh, "A Review of Multimodal Vision–Language Models: Foundations, Applications, and Future Directions," *Preprints*, 2025. Available from: https://doi.org/10.20944/preprints202510.2511.v1.
[17] Singh and T. Banerjee, "AgroMM-GSF++: Confidence-Adaptive Lightweight Multimodal Fusion for Plant Disease Recognition with Repeated Multi-Seed Cross-Validation and External Benchmarking," *Research Square Preprint*, Apr. 2026. Available from: https://doi.org/10.21203/rs.3.rs-9556788/v1.
[18] T. Banerjee, Piyush, Mukthikka V et al., "Multimodal and Large Language Model Approaches in Cybersecurity: A Systematic Review," *Research Square Preprint*, Apr. 2026. Available from: https://doi.org/10.21203/rs.3.rs-9527568/v1.
[19] T. Banerjee, Piyush NA, and Mukthikka V et al., "Multi-Modal and CNN-Based Approaches in Cybersecurity: A Comprehensive Review," *ScienceOpen Preprints*, 2026. Available from: https://doi.org/10.14293/PR2199.003446.v1.
[20] Singh and P. Mundada, "SignFuse: A Proposed Dual-Stream Cross-Modal Framework for Gloss-Free Sign Language Translation with Large Language Models," *Preprints*, 2026. Available from: https://doi.org/10.20944/preprints202604.1065.v1.
[21] Singh, S. Singh, N. Rehmani, P. Kumari, and S. Vedha Varshini, "The Role of Data Analytics in Driving Business Innovation and Economic Growth—A Comparative Study Across Industries," *International Journal of Innovative Research in Engineering and Management*, vol. 11, no. 4, pp. 33–38, 2024. Available from: https://doi.org/10.55524/ijirem.2024.11.4.5.
[22] G. Singh, T. Banerjee, and N. Ghosh, "Tracing the Evolution of Artificial Intelligence: A Review of Tools, Frameworks, and Technologies (1950–2025)," *Preprints*, 2025. Available from: https://doi.org/10.20944/preprints202511.0637.v1.

ABOUT THE AUTHORS

Piyush Thapliyal Undergraduate student (BA Programme) at the University of Delhi, School of Open Learning. Engaged in multidisciplinary research work across animation, language studies, and related interdisciplinary fields, with a focus on interdisciplinary and creative research approaches.

Purva Mundada is currently an MBA student at JSPM University, Pune, India specializing in Human Resource Management . Engaged in multidisciplinary research work across artificial intelligence, consumer behavior, digital HR, health-conscious food innovation, and cross-cultural studies.

Mukthikka V. is currently a student pursuing Aerospace Engineering at Bharath Institute of Higher Education and Research. She is involved in a CanSat project team for a national competition and has contributed to AI-related research. Her interests include artificial intelligence, computer vision, aerodynamics, propulsion systems, and sustainable environmental solutions.

Gurpreet Singh is a graduate from Endicott College of International Studies, Woosong University, South Korea. He was selected as a foundation scholar among top 30 students to visit Japan as a visiting scholar. He is currently working cross-modality and have published works in preprints. He is actively following conferences such as **ICLR, CVPR, AAI** etc.