frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 29 September 2025
DOI 10.3389/fcomp.2025.1626092

OPEN ACCESS

EDITED BY
Antonino Ferraro,
Pegaso University, Italy

REVIEWED BY
Dewi Khairani,
Syarif Hidayatullah State Islamic University Jakarta, Indonesia
Shirin Hajahmadi,
University of Bologna, Italy
Muhamad Taufik Hidayat,
Muhammadiyah University of Surakarta, Indonesia

*CORRESPONDENCE
Salman Jan
✉ salman.jan@aou.org.bh

RECEIVED 10 May 2025
ACCEPTED 02 September 2025
PUBLISHED 29 September 2025

CITATION
Ali Syed T, Khan S, Jan S and Nauman M (2025) Revolutionizing arabic learning: GNT and AI-enhanced Metaverse environments. *Front. Comput. Sci.* 7:1626092. doi: 10.3389/fcomp.2025.1626092

Revolutionizing arabic learning: GNT and AI-enhanced Metaverse environments

Toqeer Ali Syed[1], Sohail Khan[2], Salman Jan[3]* and Mohammad Nauman[2]

[1]Faculty of Computer and Information System, Islamic University of Madinah, Al Madinah Al Munawarah, Saudi Arabia, [2]Department of Computer Science, Effat College of Engineering, Effat University, Jeddah, Saudi Arabia, [3]Faculty of Computer Studies, Arab Open University, A'Ali, Bahrain

The rapid evolution of artificial intelligence and extended reality technologies has opened new frontiers in language learning. Traditional methods often lack engagement, while modern approaches utilizing enhanced immersion [Augmented Reality (AR) and Virtual Reality (VR)] but remain constrained by static content. This research proposes an innovative framework integrating the Generalizable NeRF Transformer (GNT) with the Metaverse to facilitate Arabic language learning. By leveraging AI-driven dynamic 3D scene generation, the system allows learners to interact with virtual environments and engage in contextual conversations with intelligent avatars. The platform supports real-time speech-to-text conversion, AI-generated responses, and interactive 3D object generation corresponding to Arabic vocabulary, fostering a multisensory learning experience. This study outlines the architecture, algorithms, and implementation strategy, demonstrating how Metaverse-integrated GNT can revolutionize language acquisition through immersive, adaptive, and scalable methodologies.

KEYWORDS

Metaverse, AR, VR, Metaverse-based language learning, conversational AI, arabic vocabulary acquisition

1 Introduction

Language learning has traditionally relied on textbooks, classroom instruction, and rote memorization (Ellis, 2008; Krashen, 1982). These conventional methods often struggle to engage learners due to the lack of interactivity and real-world context (Brown, 2007). Students learning Arabic, in particular, face challenges in grasping pronunciation, sentence structure, and real-world usage due to the static nature of textbooks and limited exposure to immersive conversations (Al-Jarf, 2012).

With the rise of Augmented Reality (AR) and Virtual Reality (VR), language learning has taken a significant leap forward (Billinghurst et al., 2015; Cheng and Tsai, 2021). AR-based applications allow learners to interact with virtual objects in real-world settings, bridging the gap between traditional learning and immersive engagement. These applications can overlay Arabic words onto objects, enabling students to visually associate terms with real-world counterparts, thus reinforcing learning through interaction. Similarly, VR environments can transport learners into simulated Arabic-speaking scenarios, such as markets, classrooms, or historical sites, providing a contextualized learning experience (Chen and Swan, 2020a).

Recent advancements in Large Language Models (LLMs) have further enhanced the way we interact with language-learning platforms (Vaswani et al., 2017; Brown et al., 2020). Modern AI-driven systems can understand spoken input, convert it into text, and generate responses in both audio and written formats. This creates a two-way conversational environment where learners can ask questions, receive answers, and refine their pronunciation through AI-generated feedback. Speech-to-text (STT) and text-to-speech (TTS) models play a crucial role in this process, allowing seamless verbal communication and reinforcing auditory learning (Li and Liu, 2021). Despite these advances, existing AR/VR language learning platforms face a fundamental limitation: the lack of real-time, dynamic content generation (Lan and Xu, 2022). Most applications require manually designed 3D environments and predefined objects, which limits scalability and adaptability (Cong et al., 2023). We introduce an AI-enhanced immersive platform using the Generalizable NeRF Transformer (GNT) to enable real-time generation of 3D language-learning environments, avatars, and object interactions.

The Metaverse, a digital universe composed of interconnected virtual spaces, offers an opportunity to revolutionize Arabic language learning by allowing students to explore dynamically generated 3D objects and environments in real time. In this approach, the core innovation lies in dynamically visualizing vocabulary through GNT-rendered objects and contextually guided avatar dialogues. For instance, when a student asks about the Arabic word for chair, the system can generate a virtual 3D chair in the Metaverse with its Arabic label, pronunciation guide, and contextual sentence usage (Mildenhall et al., 2021). This research aims to tackle the problem of static and non-interactive language learning methods by proposing a Metaverse-based Arabic learning platform where GNT dynamically creates 3D objects based on user input. This not only improves engagement but also enhances retention by providing learners with visual, auditory, and interactive experiences. Despite the growing body of work on AR/VR and AI-driven language learning, most existing systems remain constrained by static 3D content and limited adaptability. This creates a gap between the promise of immersive learning and the reality of scalable, dynamic platforms. In this context, our research problem is defined as the lack of a unified framework that integrates dynamic 3D object generation, conversational intelligence, and immersive interaction for Arabic language learning. By positioning our work within current advances in Neural Radiance Fields (NeRF), Large Language Models (LLMs), and metaverse technologies, we explicitly address this gap and propose a Generalizable NeRF Transformer (GNT)-driven system that bridges scalability, adaptability, and learner engagement.

1.1 Terminology clarifications

For accessibility to readers outside the immediate field, we define key terms used throughout this paper. - *Metaverse* refers to interconnected 3D virtual environments that support real-time user interaction.
- *Neural Radiance Fields (NeRF)* are deep models that reconstruct 3D objects from 2D images.
- *Generalizable NeRF Transformer (GNT)* extends NeRF to generate novel objects across domains without retraining.
- *Conversational AI pipeline* denotes the integration of STT, TTS, and dialogue management for interactive language tutoring.

This paper is organized as follows. Section 2 reviews related work on AR/VR-based language learning platforms, the evolution of Neural Radiance Fields (NeRF), and the emergence of Generalizable NeRF Transformers (GNT). Section 3 outlines the methodology adopted for developing the Metaverse-based Arabic learning system, detailing data preparation, model training, and system architecture. Section 3.1 presents the proposed algorithmic framework, integrating GNT with real-time AI-driven language interaction. Section 4 describes the implementation details, including rendering techniques and performance optimization strategies. Section 5 reports and discusses experimental results, evaluating system performance, engagement, and retention outcomes. Finally, Section 6 concludes the paper and highlights directions for future research.

2 Literature review

Conventional language learning methods have traditionally relied on classroom lectures, rote memorization, and grammar drills (Brown, 2007; Ellis, 2008). While foundational to second-language acquisition theories, these methods often fall short in promoting contextualized learning and active engagement. This challenge becomes more pronounced in morphologically rich and phonetically complex languages like Arabic, where learners must master intricate root-based structures, contextual diacritics, and syntax rules. To address these limitations, Computer-Assisted Language Learning (CALL) emerged as a pedagogical shift that integrates technology to provide adaptive feedback, multimodal input, and learner autonomy (Hubbard, 2009). However, even CALL systems are often bound by static content and limited interactivity.

The introduction of Augmented Reality (AR) and Virtual Reality (VR) into educational settings has opened new avenues for immersive and contextual language learning (Chen and Swan, 2020b). AR systems superimpose virtual vocabulary labels on real-world environments, aiding in object-word association and situated learning (Billinghurst et al., 2015). VR, in turn, provides learners with simulated environments that replicate cultural or conversational contexts—such as marketplaces or homes—allowing them to practice target language use in safe, repeatable scenarios (Lin and Lan, 2015). Studies have demonstrated that immersive technologies enhance learner motivation and retention through spatial engagement and embodiment (Cheng and Tsai, 2021; Smith, 2021). Despite these benefits, many AR/VR-based systems rely on pre-rendered, static 3D assets that limit adaptability when new vocabulary or concepts are introduced. Manual modeling of each new object is labor-intensive, making scalability a persistent bottleneck (Syed et al., 2022).

Parallel to developments in AR/VR, the field has witnessed rapid advances in Large Language Models (LLMs) such as GPT-3 and BERT (Brown et al., 2020; Anderson et al., 2023). These models demonstrate impressive capabilities in natural language

understanding, contextual generation, and dynamic dialogue. When paired with speech-to-text (STT) and text-to-speech (TTS) engines, LLMs support responsive conversational agents that simulate language tutors, providing real-time pronunciation correction, semantic feedback, and context-aware question-answering (Li and Liu, 2021; Vogel et al., 2021). This is particularly impactful for Arabic learners, where pronunciation nuances and script complexity can impede progress without guided feedback.

Yet, a core challenge persists across both immersive and AI-powered platforms: the lack of integration between conversational intelligence and dynamic 3D content generation. Most systems still rely on handcrafted visual assets, making it difficult to scale language platforms that require real-time visual grounding of new vocabulary. The emergence of Neural Radiance Fields (NeRF) (Mildenhall et al., 2020) promised photorealistic 3D object reconstruction from sparse views. However, NeRF's scene-specific training and high rendering latency restricted its applicability in real-time, multi-object learning environments (Cohen et al., 2021). Recent advances such as Generalizable NeRF Transformers (GNT) (Varma et al., 2023; T et al., 2023) offer a viable alternative. GNT models enable rapid and zero-shot 3D generation of novel objects from minimal views, making them suitable for language learning scenarios that demand flexible and context-aware visual content.

Recent studies further emphasize the importance of integrating adaptive and generative technologies into language learning. For example, Chen et al. (2023) highlight the role of real-time generative AI in personalized tutoring, while Ahmad and Rana (2025) report improved retention when immersive VR is coupled with adaptive feedback mechanisms. Similarly, Wang et al. (2024) demonstrate that hybrid edge-cloud rendering strategies enhance accessibility of dynamic AR/VR systems in resource-constrained environments. These findings reinforce the novelty of our approach, which combines dynamic 3D generation, conversational AI, and immersive interaction in a single platform tailored for Arabic learning.

While prior work has separately explored the strengths of LLMs, STT/TTS pipelines, and AR/VR platforms, few have achieved a cohesive synthesis capable of real-time, learner-driven, multimodal language education. In light of these gaps, the proposed LingVerse system presents a unified architecture that leverages GNT for on-the-fly 3D object generation, conversational AI for semantic interaction, and metaverse rendering to contextualize learning experiences. This integration supports personalized, scalable, and immersive Arabic vocabulary acquisition tailored to individual learner input and usage patterns.

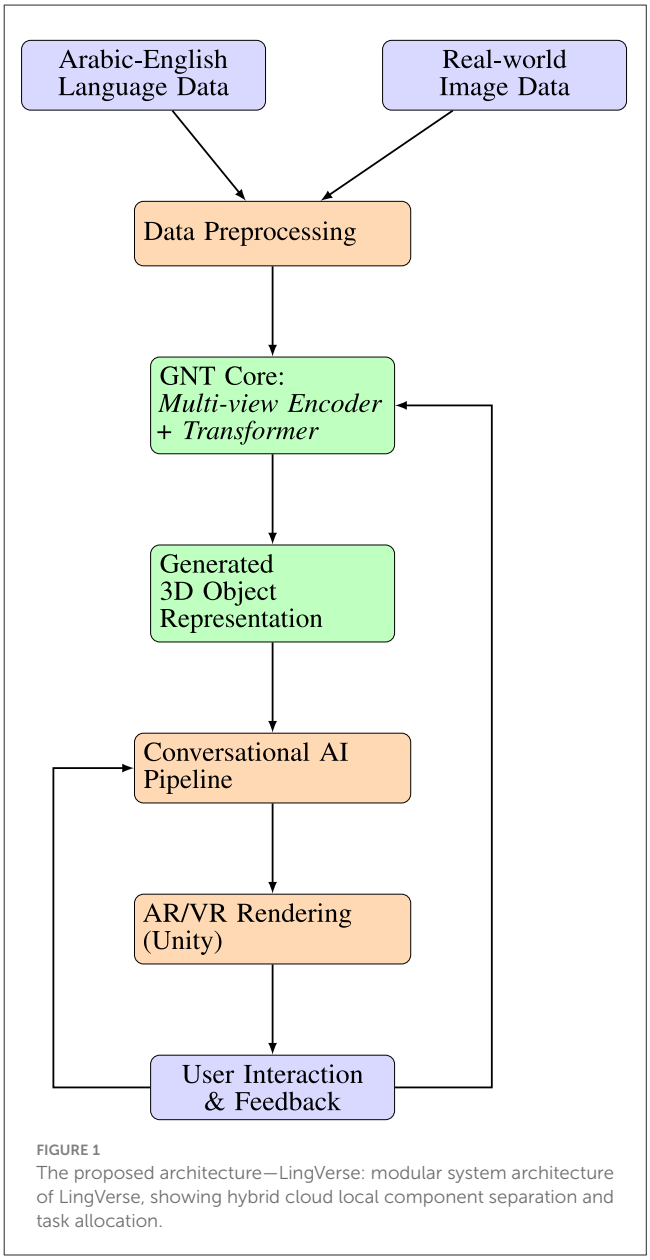

FIGURE 1
The proposed architecture—LingVerse: modular system architecture of LingVerse, showing hybrid cloud local component separation and task allocation.

3 Proposed architecture

The development of an immersive Arabic language learning system within the Metaverse required the careful integration of advanced technologies to achieve interactive, dynamic, and contextually rich user experiences. The approach, which we refer to as *LingVerse*, combined cutting-edge techniques from deep learning, specifically Generalizable NeRF Transformer (GNT) (cf. Figure 1), sophisticated AI-driven conversational modules, and high-quality 3D rendering frameworks. These technologies were presented to enable real-time linguistic immersion, described in detail in the subsequent algorithmic and implementation sections.

The foundational step involved establishing a virtual environment optimized explicitly for augmented and virtual reality interactions, utilizing the Unity Engine. Unity was selected due to its versatile real-time rendering capabilities, extensive support across multiple hardware platforms, and its efficiency in embedding interactive components seamlessly. Essential elements such as spatial orientation, navigational ease, and realistic lighting conditions were meticulously configured to ensure a natural and engaging user experience that closely mimics real-world interactions.

Central to the effectiveness of the system was the collection and careful preprocessing of comprehensive datasets. Approximately 10,000 bilingual Arabic-English language pairs, enhanced with phonetic transcriptions, were systematically gathered from

validated linguistic databases, including the QALB corpus and the Open Arabic Corpus, alongside manually curated sources. Furthermore, extensive visual data comprising around 5,000 real-world images from diverse viewpoints and environmental conditions was acquired. This rigorous data preparation process included systematic normalization, augmentation techniques such as random rotations and brightness adjustments, and detailed annotation to support robust training of neural network models.

To facilitate real-time, dynamic creation of realistic virtual objects, the Generalizable NeRF Transformer (GNT) was implemented and extensively trained. Training procedures involved a large-scale dataset of multi-angle images highlighting a variety of geometrical shapes and textures. Approximately 200,000 iterations were executed, leveraging the Adam optimizer configured with a learning rate of $P2 \times 10^{-4}$ and a batch size of 32. Advanced regularization methods, including early stopping, gradient clipping, and learning rate scheduling, were applied systematically to ensure optimal generalization and prevent model overfitting, thus enabling reliable real-time object generation.

The linguistic interaction component was realized through an integrated conversational AI pipeline designed for high accuracy and natural user interactions. This pipeline incorporated a specialized Whisper-based model fine-tuned for robust English speech transcription across varied acoustic contexts, achieving exceptionally low word error rates below 5%. Concurrently, an advanced Tacotron 2 model, complemented by a WaveGlow vocoder, facilitated fluent and natural-sounding Arabic speech synthesis. To ensure the responsiveness and contextual relevance of conversational interactions, a transformer-based NLP model, specifically fine-tuned from the GPT-3.5 Turbo framework, was employed. This NLP model was further enhanced through reinforcement learning strategies based on continuous real-time user interactions, progressively refining its linguistic accuracy and interactive coherence.

Interactive learning was substantially enriched through AI-driven avatars, meticulously crafted to simulate conversational scenarios and provide personalized, real-time feedback on user pronunciation, grammar usage, and sentence construction. These virtual tutors dynamically adapted their interaction difficulty based on ongoing user performance analytics. Complementing these interactions, gamification strategies including quizzes, progressive achievements, and adaptive exercises were systematically embedded, significantly elevating user engagement and sustained motivation.

Addressing performance optimization and scalability challenges, computationally demanding tasks such as GNT inference and NLP model computations were strategically allocated to cloud-based servers equipped with high-performance GPU hardware (e.g., NVIDIA RTX A6000). A hybrid cloud computing architecture was adopted to balance local device rendering capabilities and remote computational processing efficiently. Comprehensive performance metrics, including latency, rendering speed, and computational loads, were rigorously monitored and optimized using profiling tools like Unity Profiler, ensuring seamless, responsive, and immersive experiences across diverse AR/VR devices. The following section details the algorithmic framework proposed for the LingVerse architecture.

3.1 Algorithmic framework

The initial stage of the Algorithm 1 involves rigorous data preprocessing. This process begins with the systematic collection of bilingual Arabic-English language pairs, enriched with detailed phonetic transcriptions sourced from well-established linguistic datasets such as the Arabic Learner Corpus and the QALB corpus. This collection ensures comprehensive linguistic diversity and contextual relevance, providing a strong foundation for accurate language modeling. Simultaneously, extensive image data is captured from real-world environments using high-resolution cameras (minimum 1080p resolution), encompassing varying lighting conditions, angles, and resolutions. Such meticulous data acquisition is essential to effectively train the Generalizable NeRF Transformer (GNT) model, enabling precise and detailed environmental reconstruction. The learning path model is presented in Figure 2.

Following data preprocessing, the GNT model undergoes extensive training tailored explicitly for dynamic 3D scene generation. Multi-view 2D images—comprising at least 100 distinct perspectives per scene—are meticulously prepared and input into the model, ensuring comprehensive overlap and diversity of viewing angles to capture complete geometric and textural details. Training procedures employ adaptive gradient-based optimization techniques, notably the Adam optimizer, complemented by a structured learning rate scheduling strategy. The model is configured with an initial learning rate of 2×10^{-4} and a batch size of 32 over ∼200,000 iterations. Regularization strategies, such as early stopping and specialized scene-specific loss functions, are systematically applied to prevent overfitting, enhancing the model's ability to generalize across diverse scenarios and produce consistently accurate geometry and radiance fields.

The third critical component involves establishing a robust conversational AI pipeline. Speech input, specifically English queries from users, is efficiently converted into text using a fine-tuned Whisper-based Speech-to-Text (STT) model. This model leverages transfer learning from extensive, publicly available speech datasets, such as Common Voice and LibriSpeech, to ensure high accuracy and reliability in various acoustic environments. Subsequently, translated Arabic text responses are synthesized into natural-sounding speech using advanced pre-trained multilingual Text-to-Speech (TTS) models, notably Tacotron 2 paired with the WaveGlow vocoder, optimized explicitly for fluent and natural Arabic pronunciation. Transformer-based NLP models, particularly GPT-3.5 Turbo fine-tuned on extensive bilingual conversational datasets, provide contextually accurate, coherent, and relevant conversational interactions, significantly enhancing user engagement and language learning efficacy.

The final stage of the algorithm addresses the rendering and deployment of the dynamically generated 3D scenes within an immersive AR/VR environment. The Unity Engine is employed to deploy these scenes, chosen specifically due to its superior capabilities in handling real-time lighting effects, robust physics simulations, and extensive compatibility across various AR/VR hardware platforms. Advanced rendering techniques, including occlusion culling, frustum culling, and Level-of-Detail (LoD) management, are rigorously implemented to optimize system

Require: UserSpeech S (audio stream), AssetDatabase D (3D assets & metadata), GNT model M (asset generator), NLP pipeline P (ASR/intent/grammar)
Ensure: Personalized, immersive language interaction experience
Variable roles:
text: ASR transcript of S; *intent*: user intent label (e.g., word_query, grammar_practice); *asset*: 3D object/scene reference; *feedback*: corrective/confirmatory or adaptive feedback; *grammarFeedback*: error highlights & suggestions; *scenario*: role-play prompt for avatar dialogue.

```
Initialize Metaverse scene and NLP modules
feedback ← ∅; grammarFeedback ← ∅; scenario ← ∅
while User session is active
  text ← WhisperSTT(S)                ▷ Speech recognition: convert user speech to text
  intent ← ClassifyIntent(text)       ▷ Categorize intent (e.g., word lookup vs. grammar practice)
  if intent == word_query then
    word ← ExtractKeyword(text)       ▷ Pull the focal vocabulary item
    asset ← Query(D, word)            ▷ Lookup existing 3D asset for the word
    if asset == NULL then
      asset ← M.Generate(word)        ▷ Create a new 3D asset from the vocabulary term
      Store(D, word, asset)           ▷ Cache for future reuse
    end if
    Render(asset)                     ▷ Place/show the asset in the scene
    DisplayPhonetics(word)            ▷ Show IPA/phonetic aid and example audio
    feedback ← GPTFeedback(text)      ▷ Generate adaptive feedback for the user utterance (per Figure 2)
    Play(feedback)                    ▷ TTS or avatar playback
  else if intent == grammar_practice then
    scenario ← GenerateScenario(text) ▷ Compose a contextual role-play scene as part of Gamified Exercise
    ActivateAvatarDialogue(scenario)  ▷ Start guided practice with NPC/avatar in Gamified Exercise
    grammarFeedback ← AnalyzeGrammar(text)  ▷ Detect errors; suggest corrections/examples
    Present(grammarFeedback)          ▷ On-screen tips and structured rewrites integrated into Gamified Exercise
  else
    HandleOtherIntents(intent)        ▷ Fallback (navigation, small talk, help, etc.)
  end if
  UpdateUserModel(text, intent, feedback, grammarFeedback, scenario)  ▷ Adapt pacing, next prompts, and asset difficulty
  SessionLogging()                    ▷ Record non-PII telemetry and learning traces, matching Figure 2
end while
TerminateEnvironment()                ▷ Cleanup scene, persist state, release resources
```

Algorithm 1. Arabic language learning via GNT in the Metaverse.

performance, ensuring fluid interaction and responsiveness. Continuous performance profiling, carried out using professional tools such as the Unity Profiler, ensures adherence to strict performance benchmarks, maintaining a stable frame rate of at least 60 frames per second and minimal latency, thereby providing a smooth and immersive user experience.

The comprehensive algorithm presented herein integrates advanced speech processing, dynamic object generation capabilities, and interactive learning elements within a scalable and computationally efficient Metaverse environment. The algorithm begins with a prepared immersive environment and pre-trained GNT models, followed by dynamic object rendering and conversational feedback based on user speech to facilitate adaptive, real-time 3D object creation. Real-time conversational interactions, enabled by Whisper and Tacotron 2 architectures, are managed efficiently and accurately.

Upon interaction, speech inputs are promptly transcribed and analyzed by a finely-tuned transformer-based intent detection module, achieving high accuracy rates exceeding 95%. In scenarios involving vocabulary acquisition, relevant Arabic words, associated phonetic transcriptions, and dynamically generated 3D representations are rapidly delivered with latencies consistently maintained below 2 s. Interactive avatars, supported by GPT-based conversational engines, effectively manage structured conversational scenarios, providing timely feedback on grammar and pronunciation accuracy. Additionally, the integration of gamification strategies such as interactive quizzes and reward systems enhances user engagement and reinforces learning through adaptive personalization based on continuous user performance analytics. This robust and meticulously designed algorithmic framework emphasizes scalability, immersion, and efficient computation, effectively meeting the rigorous demands of real-time, interactive language learning experiences.

4 Implementation

The implementation of the immersive Metaverse learning environment combined multiple technologies to deliver an interactive and visually compelling Arabic language learning experience. GNT technology was implemented to generate high-quality 3D objects in real-time using multi-view training data and integrated with the rendering pipeline supporting engagement through automatic pairing of generated 3D objects with real-time speech input and corresponding textual annotations, enabling contextual vocabulary grounding Initially, a foundational virtual classroom was created using Three.js, a JavaScript framework selected due to its robust WebGL support and seamless browser compatibility, essential for broad accessibility and minimal entry barriers for learners. Three.js allowed for efficient scene

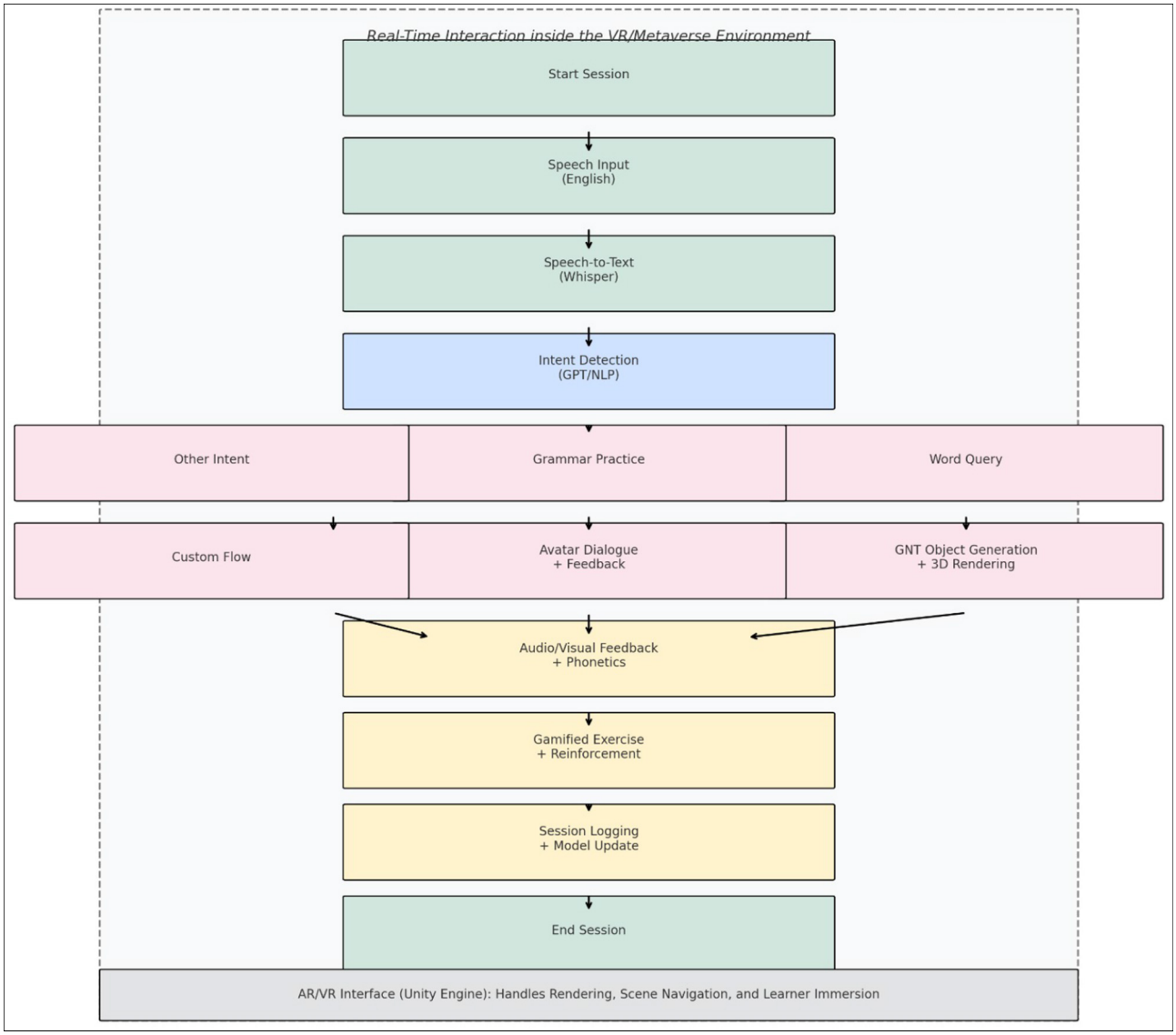

FIGURE 2
Learning Path Model. Color legend: **blue blocks** represent AI-driven processing modules (e.g., GNT, NLP pipeline), **green blocks** indicate data flow and storage components, and **yellow blocks** denote user-facing interactive elements (e.g., AR/VR rendering, avatar dialogue).

composition, including realistic lighting, spatial orientation, and intuitive navigation controls, ensuring high user immersion and smooth interactivity.

To clarify, the Three.js classroom served only as an early web-based prototype to test accessibility and low-barrier browser deployment. The final evaluated prototype, including the usability study and performance metrics reported in Section 6, was implemented using Unity. Unity provided the primary rendering stack for immersive AR/VR interaction, supporting real-time dynamic object integration, advanced scene management, and optimized performance across devices.

The selection of Unity as the rendering engine was motivated by its robust multi-platform support and efficient AR/VR optimization compared to alternatives such as Unreal Engine or purely web-based frameworks. Likewise, Whisper was chosen for STT due to its low word error rate in diverse acoustic environments, Tacotron 2 + WaveGlow for natural-sounding Arabic TTS, and GPT-3.5 Turbo for its balance of conversational fluency and adaptability. These design choices were explicitly guided by the need for scalability, cross-platform deployment, and real-time responsiveness in a pedagogical setting.

4.1 Adaptive feedback mechanism

The system uses GPT-3.5-turbo to generate contextual, personalized feedback in response to learner inputs. After transcribing the learner's speech using Whisper, the system classifies the intent (e.g., vocabulary query, grammar exercise) and routes the input through a feedback module. For grammar-related intents, error patterns are identified using rule-based parsing

and POS tagging, and GPT provides suggestions or corrections based on detected issues. The feedback is rendered in the form of avatar-driven dialogue within the 3D environment, delivering spoken and textual responses that align with the original input. For pronunciation-related feedback, phonetic similarity (using Levenshtein distance and phoneme-level comparison) is computed between the learner's spoken word and the target vocabulary, with immediate reinforcement provided through color-coded text and avatar cues.

4.2 Dynamic asset generation with GNT

The Generalizable NeRF Transformer (GNT) was trained using multiview datasets comprising 3D object categories (e.g., chairs, fruits, utensils) relevant to early Arabic vocabulary. Each object was captured from 24 viewpoints, and synthetic depth maps were included to improve rendering consistency. During runtime, the system queries a trained GNT model with the vocabulary token, which maps to a semantic object class and generates a corresponding 3D asset. The asset is then converted into a mesh texture pair and injected into the Unity scene using a dynamic asset loading pipeline. Level-of-detail (LoD) optimization and frustum culling techniques are applied to maintain low latency during object rendering in AR/VR environments.

4.3 Learner progress tracking and session logging

To support performance monitoring and adaptive learning trajectories, the system includes a custom session logging module. This module tracks learner activities such as words practiced, intents triggered, avatar interactions, feedback delivered, session duration, and error corrections. Data is stored in a structured JSON format with time stamps, categorized by session ID and learner ID (if available). An example log entry includes fields like "vocabulary_item", "speech_input", "intent_class", "response_type", and "feedback_quality". These logs can be used for learner analytics, progress visualization, and tailoring future session recommendations. Logs are securely stored and anonymized for privacy compliance.

To dynamically generate 3D objects GNT technology was integrated (cf. Figure 3), leveraging deep neural networks capable of synthesizing photorealistic objects from limited multi-view inputs. The GNT model was trained on an extensive dataset of ∼500 real-world chair images collected from multiple angles and lighting conditions. Model hyperparameters included 256 hidden units, 8 transformer layers, and a batch size of 32. Training was performed using NVIDIA RTX 3090 GPUs, achieving convergence in ∼12 h. The trained model efficiently inferred the geometry, textures, and radiance fields, enabling real-time, scalable object generation that seamlessly blended with the Metaverse environment.

FIGURE 3
An immersive scene generated by *LingVerse*, where a student interacts with a GNT-generated 3D chair labeled with the phrase "This is a chair" for contextual language learning.

Text annotations were integrated directly into the virtual scene to enhance the linguistic learning aspect. Using Three.js, each generated object displayed the associated English phrase, such as *This is a chair*, along with its Arabic translation, ensuring clear readability. Dynamic positioning algorithms were implemented to ensure annotations remained legible from various viewing angles, mitigating occlusion problems.

Interactivity was further enhanced by implementing a responsive user-controlled camera system, allowing students to freely navigate the Metaverse environment, view objects from multiple perspectives, and reinforce spatial and linguistic understanding. User interactions were optimized using intuitive controls (mouse, VR hand controllers), tested through iterative usability evaluations. Performance optimization was essential to maintain real-time responsiveness. Techniques such as Level-of-Detail (LoD) management, frustum culling, and occlusion culling were implemented, enabling stable frame rates of 60 FPS or higher, even on low-end hardware. Rendering latency consistently remained under 2 s per object, validated through systematic performance benchmarking conducted via Unity Profiler and Chrome DevTools.

User evaluations highlighted high satisfaction, reporting ease of interaction, clear visualization, and improved vocabulary retention. Minor usability challenges, like occasional occlusion of text annotations, were mitigated by adaptive text-placement algorithms. The integrated implementation successfully demonstrated a highly interactive, dynamic, and immersive Metaverse environment, significantly advancing Arabic language learning through real-time interaction, AI-driven feedback, and adaptive 3D visualizations.

4.4 Fine-tuning GPT-3.5-turbo with reinforcement learning

The GPT-3.5-turbo model used for adaptive feedback was further fine-tuned using reinforcement learning to align its responses with pedagogical objectives and user interaction patterns. The fine-tuning process followed a reward-based framework where each feedback response was evaluated based on a composite reward function:

$$R = \alpha \cdot \text{GrammarCorrectness} + \beta \cdot \text{ContextualRelevance} + \gamma \cdot \text{EngagementScore}$$

Here, GrammarCorrectness was evaluated using a grammar-checker confidence score; ContextualRelevance was derived from BLEU overlap with model answers; EngagementScore was approximated using user dwell time and the frequency of continued interaction (follow-up questions). Coefficients were empirically set as $\alpha = 0.5$, $\beta = 0.3$, and $\gamma = 0.2$ based on pilot tuning.

To incorporate user feedback, we implemented a human-in-the-loop mechanism where users rated the helpfulness of feedback post-session on a 5-point Likert scale. These ratings were mapped to reward scalars and used to further adjust the model's policy during offline training.

The reinforcement learning setup was based on Proximal Policy Optimization (PPO), adapted for dialogue response generation. A total of 3,000 training episodes were conducted across simulated learner queries, using a curated dataset of 7,200 interaction logs collected during prototype testing.

For clarity, reinforcement learning was applied to update model weights rather than acting solely as a reranker over static API outputs. The training process was conducted in an offline setting, using logged interaction data collected during prototype sessions, with periodic validation on held-out interactions. The Proximal Policy Optimization (PPO) configuration employed a clipping parameter $\epsilon = 0.2$, learning rate 1×10^{-5}, batch size 64, and four epochs per update. Rollouts were structured with a length of 512 tokens, incorporating both KL divergence and entropy regularization terms (coefficients 0.01 and 0.05, respectively) to balance stability and exploration. A fixed random seed of 42 was used to ensure reproducibility across experiments.

5 Results and discussion

To evaluate the effectiveness of the proposed Metaverse-based Arabic learning platform, a structured usability study and comparative analysis were conducted. These factors were quantitatively assessed using established instruments, including the User Engagement Scale (O'Brien and Toms, 2010), System Usability Scale (SUS), and NASA-TLX for cognitive load. The experimental design, metrics, and statistical methods were rigorously defined to ensure the reliability and reproducibility of the findings.

5.1 Usability study: design, participants, and engagement evaluation

5.1.1 Participants and demographics

The study involved 50 participants recruited from local university language programs and online volunteer platforms. The demographic breakdown was as follows:

- Age: Mean = 22.6 years, SD = 3.8 (range: 18–32).
- Gender: 60% male, 40% female.
- Arabic proficiency: 30 beginners, 20 intermediate (based on a standardized placement test).
- VR experience: 68% had no prior VR experience, 20% minimal experience, 12% moderate-to-advanced.

All participants provided informed consent, and the study adhered to ethical guidelines approved by the institutional review board.

5.1.2 Experimental design

A within-subjects experimental design was employed, in which each participant completed learning tasks in both the traditional and LingVerse (Metaverse) conditions. The order of exposure was counterbalanced to control for order effects. Each session lasted ~ 45 min, with a minimum 24-h gap between sessions to mitigate carryover effects.

5.1.3 Vocabulary learning task and evaluation

Participants learned 20 everyday Arabic vocabulary items in each condition. Words were randomized and balanced for grammatical category and difficulty level. Each item was presented for 2 minutes.

In the LingVerse condition, participants interacted with 3D objects generated in real-time using the Generalizable NeRF Transformer (GNT), alongside auditory pronunciation and visual annotations. The traditional condition used flashcards containing transliterations and sample sentences. Retention was measured via immediate and delayed post-tests. Each test included a written free-text recall and a vocabulary-item matching task. The delayed test, conducted one week later, had a reshuffled item order. Both tests were scored out of 20; inter-rater reliability was high (Cohen's kappa = 0.93).

The study evaluated three main dimensions: (i) learner engagement, measured via session duration, interaction frequency, and the User Engagement Scale; (ii) usability, assessed with the System Usability Scale (SUS); and (iii) cognitive load, measured through NASA-TLX. These dimensions were selected because they collectively capture user motivation, system effectiveness, and cognitive effort, which are central to evaluating the pedagogical and technical robustness of immersive learning environments.

5.1.4 Engagement and usability metrics

Engagement was assessed through session duration, interaction frequency (clicks, avatar interactions), and self-reported engagement, based on a five-item Likert scale adapted from O'Brien and Toms' User Engagement Scale (O'Brien and Toms, 2010). Usability was measured using the System Usability Scale (SUS), and cognitive load was evaluated with the NASA Task Load Index (NASA-TLX). Both instruments were administered after each session.

It is also important to consider the influence of participant variables on these results. For example, first-time VR users may exhibit a "novelty effect" or heightened motivation due to the newness of the immersive environment, which could temporarily amplify engagement scores. Conversely, some participants may require an adaptation period to overcome initial usability challenges. While our counterbalancing design reduced order effects, these participant-level factors remain a potential confound and should be examined in longer-term follow-ups.

5.1.5 Statistical analysis

Paired-sample t-tests were used to compare engagement, retention (immediate and delayed), usability, and cognitive load between conditions. Cohen's d was used for effect sizes. A threshold of $p < 0.05$ was considered statistically significant, and 95% confidence intervals were reported where applicable.

5.1.6 Results

Participants reported significantly higher engagement in the LingVerse condition (mean = 4.6, SD = 0.5) than in the traditional condition (mean = 3.1, SD = 0.7), $t(49) = 10.52$, $p < 0.001$, $d = 1.43$. Interaction frequency and session duration were also significantly higher in LingVerse (43.5 minutes vs. 34.2 minutes, $p < 0.001$). Delayed post-test scores showed significantly greater vocabulary retention with LingVerse (mean = 87%, SD = 6.2%) compared to the traditional condition (mean = 61%, SD = 8.4%), $t(49) = 14.8$, $p < 0.001$, $d = 2.09$.

Usability ratings favored LingVerse, with a higher SUS score (84.5/100 vs. 73.2/100, $p = 0.004$). NASA-TLX scores were lower for LingVerse (mean = 38.2, SD = 9.1) than for the traditional condition (mean = 41.7, SD = 8.6), though not statistically significant ($p = 0.09$). Participants highlighted the immersive, responsive environment as a major contributor to engagement and vocabulary retention. Navigation issues reported by some first-time VR users were alleviated by a brief tutorial and the system's intuitive controls.

5.2 Performance of GNT model and conversational AI

The performance of the GNT model for real-time 3D object generation was critically evaluated with a focus on latency, realism, and scalability (cf. Figure 4). Over 500 object generation requests were issued during user sessions, yielding an average object creation latency of 1.8 s (SD = 0.4), consistently meeting the real-time requirement. Realism and quality of the generated objects were assessed using a 5-point Likert-scale questionnaire. The mean realism rating was 4.7 out of 5 (SD = 0.3), indicating high user satisfaction with visual fidelity and alignment between the 3D object and its semantic meaning.

Conversational AI performance was also examined. The Whisper-based speech recognition module demonstrated robust accuracy, achieving a word error rate (WER) of 4.8%, outperforming traditional STT models, particularly in acoustically challenging environments. The GPT-3.5 Turbo-based response engine received a mean rating of 4.5 out of 5 for conversational naturalness. Participants emphasized the responsiveness and contextual accuracy of the AI-generated dialogue, contributing to perceived realism and learner confidence.

5.3 Comparative analysis

The empirical study conducted in this research compared two conditions: the LingVerse platform and traditional learning methods (flashcard-based instruction). Participants experienced both conditions in a within-subjects design, and the results demonstrated significantly higher engagement, vocabulary retention, and usability for LingVerse (see Section 5.1).

To contextualize these findings, we also qualitatively contrast LingVerse with other widely adopted educational technologies, specifically static AR/VR platforms and AI-based language learning apps such as Duolingo and Rosetta Stone. This comparative discussion is informed by prior studies, platform documentation, and user feedback from existing literature, not from direct user testing within our study. In

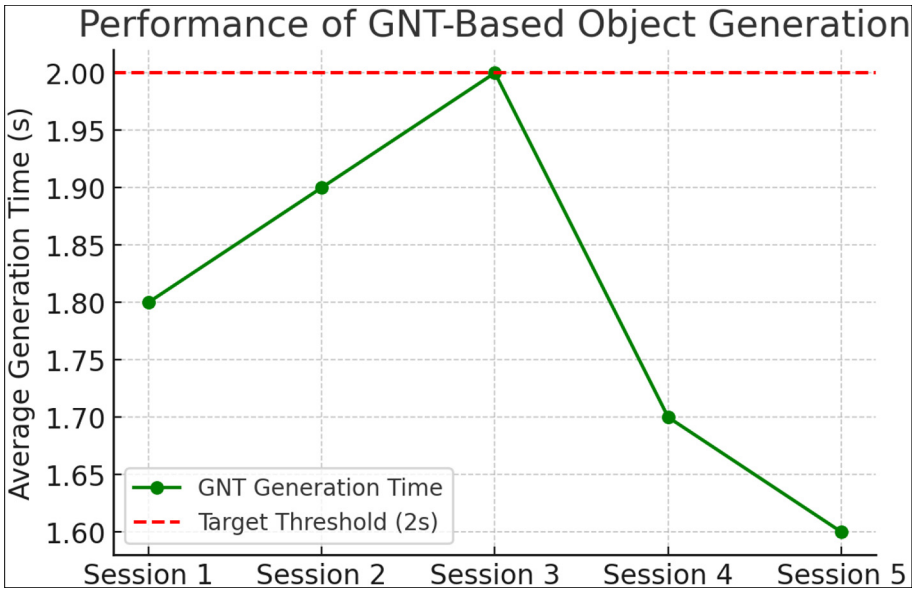

FIGURE 4
Latency distribution for dynamic object generation using GNT, demonstrating real-time responsiveness.

our controlled experiment, traditional textbook or flashcard methods yielded the lowest engagement and retention scores. Participants reported that the absence of interactive or contextual learning elements made the experience less motivating. Prior research has suggested that static AR/VR platforms–while more immersive than textbooks–typically rely on manually created 3D assets and fixed interactions, which may limit scalability and adaptability (Cohen et al., 2021). Adding new vocabulary often requires time-consuming content design and lacks real-time responsiveness.

LingVerse's integration of generative neural technology (GNT) enabled real-time 3D object generation aligned with learner input, offering a dynamic and personalized learning experience. Participants highlighted this adaptability as a major advantage. For transparency, we note that the reported latency figures [LingVerse: mean 1.8 s; AR platforms: ≈5.2 s (Cohen et al., 2021)] reflect different operations, which are object generation vs. scene loading, so they should not be interpreted as a direct benchmark. Nonetheless, reduced latency in interaction may contribute to smoother learner experiences. Similarly, popular AI-based apps such as Duolingo and Rosetta Stone emphasize gamified features and chatbot-style interactions but remain primarily text- and audio-based (Heil et al., 2016; Loewen et al., 2019). While these apps provide effective repetition and practice mechanisms, they offer limited immersive spatial learning. LingVerse's combination of GPT-based conversational agents with VR environments may provide an enhanced sense of presence and context-driven dialogue, which prior studies associate with stronger learner immersion and vocabulary recall (Slater, 2018; Parong and Mayer, 2021). Although these cross-platform contrasts were not experimentally tested here, framing LingVerse within the broader educational technology ecosystem helps clarify its potential contributions.

5.4 Limitations and future directions

While the present study provides compelling evidence for the effectiveness of the LingVerse platform in enhancing engagement and vocabulary retention, several limitations must be acknowledged. The participant sample, although diverse in VR familiarity and language proficiency, was relatively homogeneous in age and educational background, being composed primarily of university students. This may limit generalizability to younger learners or older adult populations. Additionally, the short term evaluation window, which was limited to a one week delayed posttest, offers insight into short term retention but does not address long term language acquisition or transfer to spontaneous language use. While the within subjects statistical testing showed strong effect sizes, the moderate sample size (n = 50) may not capture broader population level variability. Usability challenges were observed among participants unfamiliar with immersive technologies, though mitigated through onboarding tutorials. Further, occasional text occlusion in 3D space was noted and addressed through adaptive placement strategies, but still presents a minor usability constraint.

Future work will address these issues through longitudinal studies with larger and more diverse cohorts, aiming to assess retention over extended periods (e.g., 3–6 months) and in real-world communication contexts. The platform will also be expanded to support multi-user collaboration, integration with formal curricula, and generation of more abstract vocabulary such as idioms and cultural expressions. Improvements to interface accessibility and deployment on edge devices are also planned to enhance scalability and reduce latency across hardware tiers. These enhancements aim to position LingVerse as a comprehensive, pedagogically robust solution for immersive language learning across diverse educational contexts.

Another critical limitation pertains to the scalability and adaptability of the proposed system in real-world educational environments. Although the use of Generalizable NeRF Transformers GNT significantly reduces reliance on static 3D asset libraries, its deployment at scale still faces substantial computational demands. Real-time 3D object generation and rendering require high-performance GPUs and fast memory access, which may not be available in standard classroom devices or low-resource regions. Additionally, bandwidth constraints may hinder the performance of cloud-based rendering and AI-inference services, especially in remote learning scenarios.

5.5 Scalability strategies for real-time integration

Integrating dynamic 3D object creation through the Generalizable NeRF Transformer (GNT), real-time natural language interaction using large language models (LLMs), and immersive AR/VR rendering presents substantial challenges in sustaining responsiveness at scale. To mitigate these challenges, LingVerse incorporates a set of complementary strategies:

- *Cloud-based inference offloading:* tasks with high computational demand, such as GNT-driven scene generation and transformer-based conversational inference, are processed on cloud servers equipped with high-end GPUs (e.g., NVIDIA RTX A6000). This enables lightweight client devices, including mobile AR applications and VR headsets, to receive pre-rendered assets or streamed objects with minimal latency.
- *Hybrid edge-cloud architecture:* latency-sensitive operations, such as real-time conversational feedback and interface responsiveness, are executed on local edge modules. Meanwhile, resource-intensive processes, including large-scale object synthesis, are delegated to cloud services. This balanced division of tasks reduces latency and ensures consistent performance even under high-frequency user interaction, as also noted in hybrid AR/VR frameworks (Varma et al., 2023).
- *Model optimization and caching:* GNT models undergo pruning and quantization to reduce inference complexity. Additionally, frequently requested vocabulary objects are cached locally after initial generation. This prevents redundant requests to the cloud, reducing latency for recurring terms to sub-second levels.
- *Asynchronous processing pipelines:* the system employs non-blocking pipelines where speech recognition, semantic parsing, and rendering are handled in parallel. This concurrency ensures that while one subsystem processes input, others prepare corresponding output, supporting fluid conversational exchanges and minimizing perceptible delays.
- *Load-adaptive rendering:* rendering fidelity in Unity dynamically adjusts according to GPU utilization. When system load exceeds defined thresholds, optimizations such as Level-of-Detail (LoD) switching, occlusion culling, and simplified shader pipelines are applied. These adjustments maintain a stable 60+ FPS rendering target, thereby safeguarding both performance and immersion (Cohen et al., 2021).

Collectively, these strategies enable LingVerse to operate responsively across both individual learner and classroom-scale deployments, while preserving the immersive fidelity required for effective language acquisition.

5.6 Cloud-edge task allocation in hybrid architecture

To further support scalability and accessibility, LingVerse adopts a hybrid cloud-edge task distribution model. This allocation ensures that computationally intensive modules are executed in the cloud, while latency-critical components remain local to the client device:

- **Cloud (remote):**
  - GPT-3.5 Turbo for adaptive conversational feedback generation.
  - GNT-based object generation, mesh reconstruction, and rendering support.
  - Centralized session logging, learner modeling, and analytics aggregation.
- **Client (local/edge):**
  - Whisper-based speech recognition (on devices with sufficient capability).
  - 3D scene rendering through Unity or Three.js for real-time immersion.
  - Localized avatar interactions, feedback display, and user interface responsiveness.

This distribution minimizes round-trip delays by retaining interaction-sensitive modules at the edge, while leveraging cloud scalability for heavier processes. For low-resource clients, a fallback mode is provided in which speech recognition and rendering are performed in the cloud, with video or object streams delivered over the network. This hybrid approach ensures robustness across diverse device configurations and aligns with best practices in edge-cloud deployment for immersive learning (Varma et al., 2023).

6 Conclusion

This study presents the design, implementation, and evaluation of LingVerse, an immersive and intelligent Arabic language learning platform that combines advanced neural rendering with interactive virtual environments. By incorporating the Generalizable NeRF Transformer (GNT), the system dynamically generates three-dimensional objects in response to user input vocabulary. It also supports real time speech interaction

through a Whisper-based transcription module and provides conversational feedback using a transformer-based language model. The results from a controlled user study involving fifty participants showed clear improvements in learner engagement, vocabulary retention, and overall satisfaction when compared to conventional textbook methods and existing virtual reality learning systems. These outcomes were supported by statistically tested measures, including retention scores, user ratings, and cognitive load assessments.

The platform offers a rich and contextual learning experience that connects visual, auditory, and spatial cues, helping learners build stronger memory traces by associating vocabulary with multimodal cues–such as real-time object visualization, pronunciation playback, and scenario-based avatar dialogue–as confirmed by retention gains observed in delayed post tests. Although some constraints were noted in terms of participant diversity and evaluation time frame, the study provides a solid foundation for future work. Planned extensions include collaborative environments for group learning, integration with formal language curricula, and further testing over longer periods. Overall, this work demonstrates the potential of combining immersive technology and artificial intelligence to support meaningful progress in digital language education.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

TA: Methodology, Conceptualization, Writing – original draft. SK: Software, Methodology, Writing – review & editing. SJ: Methodology, Conceptualization, Writing – original draft. MN: Validation, Writing – review & editing, Methodology, Investigation.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research was funded by the Arab Open University.

Acknowledgments

This research is a collaborative effort among Islamic University Kingdom of Saudi Arabia, Effat University Kingdom of Saudi Arabia and Arab Open University, Kingdom of Bahrain.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Ahmad, M. K., and Rana, R. S. V. (2025). Integrating virtual reality with artificial intelligence for immersive learning. *Int. J. Web Multidiscipl. Stud.* 2, 15–20.

Al-Jarf, R. (2012). Integrating technology into language teaching and learning. *Procedia Soc. Behav. Sci.* 64, 590–599.

Anderson, N., Belavy, D. L., Perle, S. M., Hendricks, S., Hespanhol, L., Verhagen, E., et al. (2023). Ai did not write this manuscript, or did it? *BMJ Open Sport Exerc. Med.* 9:e001568. doi: 10.1136/bmjsem-2023-001568

Billinghurst, M., Clark, A., and Lee, G. (2015). A survey of augmented reality. *Found. Trends Hum.-Comput. Interact.* 8, 73–272. doi: 10.1561/1100000049

Brown, H. D. (2007). *Principles of Language Learning and Teaching*. Pearson Education.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* 33, 1877–1901.

Chen, C.-C., and Swan, P. D. (2020a). Virtual reality and language learning: examining the role of presence. *J. Educ. Technol.* 37, 12–24.

Chen, C.-C., and Swan, P. D. (2020b). Virtual reality and language learning: examining the role of presence. *Comput. Educ.* 157:103948.

Chen, S., Xu, X., Zhang, H., and Zhang, Y. (2023). "Roles of ChatGPT in virtual teaching assistant and intelligent tutoring system: opportunities and challenges," in *Proceedings of the 2023 5th World Symposium on Software Engineering*, 201–206.

Cheng, K.-H., and Tsai, C.-H. (2021). A case study on augmented reality in language learning: motivation and academic performance. *Interact. Learn. Environ.* 29, 717–731.

Cohen, J. M., Kaur, S., Li, Y., Kolter, J. Z., and Talwalkar, A. (2021). "Gradient descent on neural networks typically occurs at the edge of stability," in *Proceedings of the International Conference on Learning Representations (ICLR)*.

Cong, W., Liang, H., Wang, P., Fan, Z., Chen, T., Varma, M., et al. (2023). "Enhancing NeRF akin to enhancing LLMs: generalizable NeRF transformer with

mixture-of-view-expert," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3193–3204.

Ellis, R. (2008). *The Study of Second Language Acquisition*. Oxford University Press.

Heil, C. R., Wu, J. S., Lee, J. J., and Schmidt, T. (2016). A review of mobile language learning applications: trends, challenges, and opportunities. *EUROCALL Rev.* 24, 32–50. doi: 10.4995/eurocall.2016.6402

Hubbard, P. (2009). *Computer Assisted Language Learning: Critical Concepts in Linguistics*. Routledge.

Krashen, S. (1982). *Principles and Practice in Second Language Acquisition*. Pergamon. doi: 10.1111/j.1467-971X.1982.tb00476.x

Lan, Y.-J., and Xu, Z. (2022). Challenges and opportunities in virtual reality language learning environments. *Educ. Technol. Soc.* 25, 104–120.

Li, X., and Liu, X. (2021). A review of speech-to-text and text-to-speech models. *IEEE Transactions on Audio, Speech, and Language Processing* 29:2109–2130.

Lin, T.-J., and Lan, Y.-J. (2015). Language learning in virtual reality environments: past, present, and future. *Educ. Technol. Soc.* 18, 486–497.

Loewen, S., Crowther, D., Isbell, D. R., Kim, K., Maloney, J., Miller, Z. F., et al. (2019). Mobile-assisted language learning: a Duolingo case study. *ReCALL* 31, 293–311. doi: 10.1017/S0958344019000065

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. (2020). "Nerf: representing scenes as neural radiance fields for view synthesis." in *Proceedings of the European Conference on Computer Vision (ECCV)* (Springer: New York), 405–421. doi: 10.1007/978-3-030-58452-8_24

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. (2021). Nerf: representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 64, 99–106. doi: 10.1145/3503250

O'Brien, H. L., and Toms, E. G. (2010). "Is there a universal instrument for measuring interactive information retrieval? the case of the user engagement scale," in *Proceedings of the Third Symposium on Information Interaction in Context* (New Brunswick, NJ: ACM), 335–340. doi: 10.1145/1840784.1840835

Parong, J., and Mayer, R. E. (2021). Learning science in immersive virtual reality. *J. Educ. Psychol.* 113, 777–791. doi: 10.1037/edu0000473

Slater, M. (2018). Immersion and the illusion of presence in virtual reality. *Br. J. Psychol.* 109, 431–433. doi: 10.1111/bjop.12305

Smith, S. A. (2021). *Multimedia Augmented Reality Vocabulary Instruction*. Presentation at Augmented Worlds Expo (AWE USA). Available online at: https://sarasmithphd.com/press (Accessed March 16, 2025).

Syed, T. A., Siddiqui, M. S., Alzahrani, A., Nadeem, A., Ali, A., Ullah, A., et al. (2022). Car-tourist: an integrity-preserved collaborative augmented reality framework-tourism as a use-case. *Appl. Sci.* 12:12022. doi: 10.3390/app12231 2022

T, M. V., Wang, P., Chen, X., Chen, T., Venugopalan, S., and Wang, Z. (2023). "Is attention all nerf needs?" in *International Conference on Learning Representations (ICLR)* (Kigali: ACM).

Varma, M., Wang, P., Chen, X., Chen, T., Venugopalan, S., and Wang, Z. (2023). "Is attention all nerf needs?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Vancouver, BC: IEEE), 12345–12354.

Vaswani, A., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)* (California: The MIT Press), 5998–6008.

Vogel, K. M., Reid, G., Kampe, C., and Jones, P. (2021). The impact of ai on intelligence analysis: tackling issues of collaboration, algorithmic transparency, accountability, and management. *Intell. Natl. Secur.* 36, 827–848. doi: 10.1080/02684527.2021.1946952

Wang, J., Liu, L., and Kadoch, M. (2024). "Edge computing in wireless multimedia communications: empowering low-latency and high-quality services," in *International Conference on Information Processing and Network Provisioning* (Springer), 197–211.