User Acceptance Behavior Analysis of Multimodal Generative AI

Kuo-Tai Lu [1]*

[1]*Department of Business Administration, National Yunlin University of Science and Technology, Taiwan;

*Corresponding Author: jackylu@gmail.com

DOI: https://doi.org/10.30211/JIC.202503.008

Submitted: Apr. 19, 2025 Accepted: Jun. 10, 2025

ABSTRACT

This study explores user acceptance behavior toward multimodal generative artificial intelligence (MGAI) by integrating the technology acceptance model (TAM) and media richness theory (MRT) into a cross-theoretical analytical framework. The proposed model examines the relationships among media richness, perceived ease of use, perceived usefulness, user attitude, and continued usage intention. Empirical analysis was conducted using structural equation modeling with a sample of 419 users who have experience of MGAI technologies. The results indicate that media richness significantly enhances perceived usefulness and user attitude; however, its impact on perceived ease of use is limited, which suggests that increased familiarity with technology may reduce the importance of interface simplicity. Additionally, perceived usefulness plays a critical mediating role between attitude and continued usage intention. This study extends the theoretical application of TAM and MRT in the context of multimodal AI, and provides practical recommendations for optimizing technology design and user experience.

Keywords: Multimodal generative AI, technology acceptance model, media richness theory.

1. Introduction

With the rapid advancement of artificial intelligence (AI) technologies, generative AI has emerged as a core driver of industrial innovation and transformation. Since OpenAI's launch of ChatGPT at the end of 2022, generative AI has rapidly expanded into multimodal applications, including text, image, voice, and video; this marks the beginning of the practical and commercial phase of multimodal generative AI (MGAI). Unlike traditional unimodal AI, MGAI is capable of integrating various data types—such as text, image, voice, and video—to provide comprehensive and context-aware intelligent services, which significantly enhance the efficiency and quality of human–computer interaction [1].

MGAI technologies have been widely applied across medical diagnostics [2], educational support [3], smart home solutions [4], and e-commerce, thus demonstrating the diverse value of cross-modal integration. In medical settings, diagnostic systems that combine voice descriptions with medical imaging can provide more accurate clinical recommendations and personalized treatment

plans [5]. In educational applications, MGAI leverages speech recognition, natural language understanding, and visual perception technologies to substantially improve learning interactions and content absorption rates [6]. Furthermore, AI assistants, intelligent customer service, fitness applications, and smart home devices are increasingly relying on multimodal technologies to meet users' diverse interaction needs [7].

Despite the growing ubiquity of MGAI applications, challenges remain regarding users' acceptance and continued usage intention (CUI). Previous studies have indicated that users' CUI for emerging technologies is primarily influenced by their perceived usefulness (PU) and perceived ease of use (PEOU) [8, 9]. However, while the traditional technology acceptance model (TAM) explains technology acceptance behavior, its assumption that the impact of PU and PEOU on attitudes is linear may not fully capture the complex acceptance process involved in multimodal interactions with MGAI. The high media richness (MR) of MGAI systems—including real-time feedback and multi-channel information integration—not only affects users'' functional perceptions, but also engages emotional connections and immersive experiences. These psychological mechanisms extend beyond the traditional scope of TAM. Media richness theory (MRT) posits that the richness of a medium (e.g., diversity, immediacy) significantly enhances information delivery effectiveness and user engagement [10, 11], and thus offers a complementary perspective for understanding MGAI acceptance behavior.

Although research integrating TAM and MRT has gradually gained attention, existing literature still lacks a systematic exploration of how the media characteristics of MGAI affect users' attitudes and CUI. The complex interfaces and high information density of MGAI may alter users' evaluative logic regarding PU and PEOU, and thus render the single TAM model insufficient for explaining this dynamic process. Therefore, this study integrates TAM and MRT to propose a cross-theoretical framework that aims to reveal the psychological mechanisms between multimodal media characteristics and user acceptance behaviors; this will fill the theoretical gap in MGAI adoption research.

2. Literature Review

2.1 Development and Applications of MGAI

MGAI represents a significant breakthrough in the evolution of generative AI from unimodal to cross-modal integration. By simultaneously processing voice, images, text, and video, MGAI substantially enhances the naturalness of human–computer interactions and contextual awareness [12, 13]. MGAI leverages large language models (LLMs) such as GPT-4, visual generation models including DALL·E 3 and Stable Diffusion, as well as voice synthesis technologies (e.g., VALL-E), to achieve information integration through cross-modal pre-training [14]. Recent models, such as NExT-GPT and LLaVA, have further advanced cross-modal learning, by significantly boosting AI's contextual understanding and generative capabilities in complex tasks [13, 15].

The technological advancements in MGAI stem from three key factors: First, contrastive learning in CLIP has enhanced semantic alignment between images and text [12]. Second, the Transformer architecture enables unified modeling of multimodal data, with models such as Flamingo

substantially improving image description and visual question-answering capabilities [16]. For example, recent studies have proposed multimodal fusion models integrating Faster R-CNN visual features with BERT semantic embeddings, employing ranking-based hybrid training strategies to enhance performance on complex VQA tasks, significantly outperforming state-of-the-art methods [17]. Lastly, generative adversarial networks (GANs) and diffusion models have optimized high-quality image and video generation; they have found extensive applications in creative design and virtual reality [18]. These technologies not only process multimodal inputs, but also generate high-fidelity and context-aware outputs; this significantly enhances the realism and user acceptance of human–computer interactions [19].

MGAI has been widely applied across medicine, education, creative industries, and commercial domains, which showcases its unique advantages in cross-modal integration. In the medical field, MGAI integrates voice, image, and text data to achieve precise diagnostics and personalized treatment. For example, [2] demonstrated that MGAI systems combining voice medical histories with CT imaging can automatically generate diagnostic reports, thereby improving clinical decision-making efficiency. NExT-GPT further enables multimodal medical dialogues, through integrating voice and image data to provide real-time diagnostic recommendations [13].

In education, MGAI supports personalized learning and interactive teaching. According to [6], MGAI platforms that combine speech recognition and visual content generate adaptive learning materials based on student needs; this significantly enhances learning engagement and outcomes. For instance, LLaVA supports interactive question-and-answer sessions with images and text, thus promoting inquiry-based learning in scientific courses [20]. Relatedly, studies on student motivation reveal that higher learning motivation positively predicts flow experience, which in turn enhances learning satisfaction, underscoring the importance of immersive and engaging environments in educational technology [21].

In the creative industry, MGAI has become a vital tool for digital content creation. [18] illustrated the application of diffusion models in high-resolution image generation, as it effectively shortens the animation design and scene modeling cycle. Currently, models such as Stable Diffusion and MidJourney generate high-quality visual content from text prompts, while Runway's Gen-2 and VALL-E simplify short video production and virtual dubbing. In the commercial domain, MGAI optimizes consumer experiences through multimodal interactions. For example, [19] analyzed consumer evaluations of unimodal versus multimodal voice assistants, and provided fresh insights into multimodal interactions.

The development of MGAI has redefined the standards of human–computer interaction. Its real-time feedback, diverse outputs, and high contextual awareness significantly influence user cognition and emotional evaluation. According to MRT, multimodal technologies enhance communication efficiency and interactive engagement through real-time delivery of multi-channel information [10]. However, its complex interfaces may reduce PEOU [22]. According to the TAM, users' CUI is influenced by PU and PEOU [8]. However, the multimodal characteristics of MGAI may challenge the traditional assumptions of TAM; particularly in terms of social presence and emotional connection

mechanisms, which require further exploration.

2.2 Media Richness and Interactive Influence

MRT, proposed by Daft and Lengel, explains how different media have varied effectiveness in managing information uncertainty and ambiguity [10]. MRT posits that media richness is determined by four main factors: immediacy of feedback, message variety, language naturalness, and personalization. High-richness media can effectively convey complex information, thus enhancing communication efficiency and understanding [10]. With the progress of AI technologies, MRT has been widely applied to human–computer interaction research; particularly to explore how MGAI enhances users' experience and acceptance through voice, image, and text integration.

The multimodal design of MGAI exhibits high media richness, which significantly influences users' cognition, emotional connections, and behavioral intentions. Unlike unimodal AI, MGAI simulates real communication scenarios through the synergistic effects of voice, image, and text, to provide an intuitive and immersive interactive experience [19]. For instance, MGAI educational platforms that combine speech recognition and visual cues can respond to learning needs in real time, thus significantly improving learning efficiency and engagement [20]. This high richness not only reduces cognitive load, but also strengthens social presence and interactive expectations, thereby facilitating technology acceptance.

Integrating MRT with the TAM can further explain MGAI's acceptance mechanisms. MRT views media richness as an external variable that affects PU, PEOU, and attitude (ATT), and thus shapes technology adoption intentions [23]. Empirical studies support this perspective. For example, Chatterjee et al. found that AI customer service systems that combine voice and visual feedback significantly enhance perceptions of functionality and operational convenience [24]. The multimodal characteristics of MGAI not only improve information transmission efficiency, but also foster positive evaluations and CUI through immersion and emotional connection [22].

According to MRT, high-richness media can reduce information uncertainty and improve task efficiency through real-time feedback and message diversity [10]. In MGAI scenarios, the integration of voice, image, and text provides rich and context-aware information outputs, which enhance users' perceptions of the system's practical value [22]. For example, Heirati's study showed that multimodal interactive voice assistants significantly improve users' evaluations of system functionality [19]. Thus, we propose the following hypotheses:

H1: MR positively influences PU.

H2: MR positively influences ATT.

H3: MR positively influences PEOU.

2.3 Technology Acceptance Model and Cognitive Processes

The TAM, proposed by Davis, serves as a foundational framework for understanding individual acceptance of emerging technologies. TAM emphasizes that PU and PEOU are the core constructs that influence users' ATT and behavioral intentions [8]. PU reflects the extent to which technology enhances work or learning efficiency, while PEOU indicates the simplicity of operation and reduced

cognitive load [25]. Due to its simplicity and strong predictive capability, TAM has become the primary theoretical model for technology acceptance research.

Empirical studies have shown that PU and PEOU not only impact initial adoption, but are also strongly associated with CUI and satisfaction [9]. For example, Li et al. found that PU and PEOU significantly predict usage intentions for digital payment systems, with digital literacy moderating the strength of their impact [26].

Similarly, a study conducted among university students in Nepal revealed that perceived usefulness was the most influential factor in adopting cashless transaction systems, with perceived ease of use also playing a significant role. Although the context differs from multimodal AI, this study underscores the universal applicability of TAM constructs in predicting user behavior across digital systems. Such findings provide cross-contextual empirical support for using PU and PEOU to explain acceptance of emerging technologies, including MGAI [27].

Furthermore, PU and PEOU are not independent: Davis pointed out that PEOU positively influences PU [8], while Kim et al. demonstrated that in high-functionality technological contexts, PU can reciprocally enhance PEOU, as users recognize its benefits and become more willing to learn [28].

In the context of MGAI, the interaction between PU and PEOU becomes even more intricate. MGAI integrates voice, image, and text to provide high-performance task support, thus significantly enhancing PU [24]. However, while the multimodal design boosts functionality, its complexity may elevate operational barriers; this affects the formation of PEOU [19]. Therefore, within MGAI contexts, the influence of PU on PEOU becomes more crucial. If users perceive high performance, they are more willing to engage in learning, thus reducing perceptions of operational obstacles. Based on TAM, when users perceive technological benefits, they are more inclined to learn, and experience less anxiety regarding system complexity [8]. For instance, research by Chatterjee et al. showed that high-functionality AI customer service systems significantly improve operational convenience [24]. Compared to unimodal technologies, MGAI's multimodal design strengthens the impact of PU on PEOU; although this causal relationship may be moderated by interface complexity, which necessitates further verification. Hence, the following hypothesis is proposed:

H4: PU positively influences PEOU.

According to the TAM framework, PU and PEOU have significant effects on ATT toward technology use. PU strengthens users' perceptions of the technology's benefits, and thus directly promotes a positive attitude [8]; while PEOU reduces operational burdens, thereby indirectly influencing technology acceptance [25]. Empirical research supports this perspective; for example, Harnida and Mardah found that PU and PEOU significantly influence user attitudes toward electronic payment platforms [29], while Chung and Nam demonstrated that PEOU is particularly impactful for first-time users on online-to-offline platforms [30].

In the context of MGAI, the impact of PU and PEOU becomes even more pronounced. The multimodal design enhances PU through real-time feedback and personalized experiences [24]. For example, research by Lu et al. showed that the personalized features of MGAI-based educational

platforms significantly improve learning efficiency and strengthen user attitudes [28]. Simultaneously, the multimodal interface simplifies operations and enhances PEOU [19]. However, the high degree of automation and contextual reasoning in MGAI may lead users to prioritize functionality over ease of use, thereby potentially weakening the direct influence of PEOU on ATT [19].

MRT further complements the mechanism of PU and PEOU's influence. Yuan et al. pointed out that multimodal designs with high media richness enhance users' perceptions of system performance and user-friendliness, and thus promote a positive attitude [31]. MGAI's multimodal interactions, facilitated through real-time feedback and immersion, amplify the impacts of PU and PEOU on ATT [24]. Based on the theoretical and empirical insights, the following hypotheses are proposed:

H5: PEOU positively influences ATT.

H6: PU positively influences ATT.

2.4 Attitudes and Continued Usage Intention Mechanism

ATT serves as a crucial mediating construct in technology acceptance research, as it links cognitive evaluation with behavioral intentions. According to the theory of reasoned action (TRA) and the TAM, ATT represents an individual's overall evaluation of the behavioral objective (e.g., using AI technology), and thus influences continued use, recommendations, or paid upgrades [8, 29]. In the context of MGAI, the formation of ATT is influenced by the complexity and immersion of multimodal interactions. MGAI's multimodal design—integrating voice, image, and text—enhances intuitiveness and emotional connections through real-time feedback and personalized outputs, thereby strengthening positive attitudes [24]. For instance, Lu et al. indicated that the personalized learning features of MGAI educational platforms significantly improve students' positive evaluations, and thus promote long-term usage intentions [28].

CUI is a critical indicator of the long-term adoption and stability of technology use; it is widely applied in domains such as social media, educational platforms, and medical AI [25, 31]. In the MGAI context, high interactivity and low learning costs make initial positive attitudes more likely to transition into long-term usage behaviors [4]. Unlike unimodal AI, MGAI's multimodal interactions simulate human communication scenarios; this reinforces emotional recognition and trust, and thus drives CUI [19]. The theory of planned behavior (TPB) supports the connection between ATT and CUI; it emphasizes that positive attitudes are key antecedents of behavioral intentions, particularly in high-autonomy decision-making scenarios [32]. Similarly, Lee et al. demonstrated that highly interactive MGAI systems enhance user loyalty through positive attitudes, especially under contextualized feedback [4].

However, the impact of ATT on CUI is not strictly linear: it is moderated by factors such as trust, prior experience, and interaction quality. Additionally, technology familiarity and usage frequency also moderate the relationship between ATT and CUI. Heirati's research showed that experienced users are more inclined to trust MGAI, which amplifies the influence of positive attitudes; conversely, novice users may struggle with system complexity, which weakens this effect [19].

MGAI's multimodal characteristics offer new perspectives on the relationship between ATT and CUI. Unlike unimodal systems, MGAI leverages voice and visual personalized feedback to simulate

human-like interactions, which significantly enhances emotional recognition [24]. For instance, Chatterjee et al. showed that MGAI-based customer service systems that integrate speech recognition and visual cues substantially increase interaction satisfaction, thus subsequently strengthening CUI [4]. However, the moderating roles of trust and interaction quality require further exploration. Based on TAM, TPB, and empirical evidence, the following hypothesis is proposed:

H7: ATT positively influences CUI.

3. Research Design

3.1 Research Framework

This study aims to explore the impact of MGAI on user acceptance behavior and continued usage intention. Based on TAM and MRT, a cross-theoretical integrative framework is proposed. This framework incorporates the following core variables: media richness (MR), PEOU, PU, ATT, and CUI. Through a theoretical analysis of multimodal interactions, this study investigates the acceptance process and behavioral intentions of MGAI technology, by verifying the causal relationships and interactive effects among these constructs. The research framework is illustrated in Fig. 1, which presents the hypothesized paths and expected impact mechanisms among the constructs.

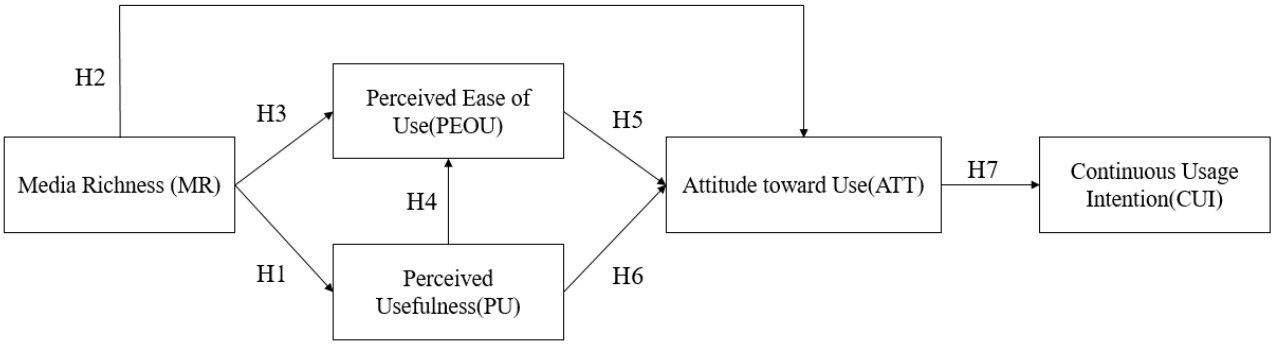

Fig 1. esearch Architecture

3.2 Research Methodology Overview

To validate the research hypotheses, this study employs structural equation modeling (SEM) for empirical analysis. SEM is a powerful multivariate statistical method capable of simultaneously analyzing multiple causal relationships, and evaluating model fit and path effects among constructs. It is particularly suitable for exploring the interactive effects of variables in complex structures.

This research uses Jamovi 2.5.5 as the primary tool for data analysis, given its comprehensive support for SEM modeling, intuitive interface, and high computational efficiency. Jamovi's open-source nature further facilitates result verification and expansion in subsequent research, to ensure the transparency and reproducibility of analysis results.

3.3 Measurement Scales

This study employed a quantitative survey method as the primary data collection tool. Prior to the survey, all participants were informed of the study's objectives and provided informed consent. The questionnaire was reviewed by experts and underwent a small-scale pilot test to ensure clarity

and content validity.

The survey instrument was divided into two sections; the first collected demographic information, including respondents' gender, age, and educational background. The second section assessed participants' behavioral responses toward MGAI services, covering the constructs of MR, PEOU, PU, ATT, and CUI.

All items were measured using a five-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree), to ensure the reliability and validity of the collected data.

3.4 Data Collection and Sampling

The constructs in this study were measured using validated scales from existing literature. MR was measured using a four-item scale [34]. A sample item is: "Multimodal AI, which integrates text, images, and audio analysis, can provide real-time feedback and support rapid interactions." The scale demonstrated high internal consistency, with a Cronbach's α of 0.904. PEOU and PU were measured using four-item scales [35]. Representative items include: "It is easy for me to complete tasks using multimodal AI" for PEOU, and "Using multimodal AI improves the quality of my work" for PU. According to previous studies, the Cronbach's α values for these scales ranged from 0.86 to 0.98 for PEOU and 0.87 to 0.98 for PU, which indicate excellent reliability. ATT was measured using a three-item scale [34]. A sample item is: "I am interested in using multimodal AI." The scale achieved a Cronbach's α of 0.946; this reflects strong internal consistency. CUI was measured with a three-item scale [36]. A representative item is: "I am willing to continue using multimodal AI." The scale demonstrated good reliability, with a Cronbach's α of 0.87.

4. Data Analysis and Results

4.1 Descriptive Statistics

Data for this study were collected through a structured online questionnaire, facilitated by the online survey platform "Meta Survey Marketing Research Co., Ltd." (https://www.drsurveydone.com/), which managed questionnaire distribution and sample administration. The survey design and implementation followed strict procedures to ensure data quality and research validity. The survey's introductory page included explanations of the study's purpose, data usage, confidentiality guarantees, and definitions and application examples of MGAI technologies, such as ChatGPT, Claude, and Google Gemini, to help participants understand the relevant concepts.

The questionnaire comprised three main sections: (1) Research Introduction and Informed Consent; (2) Demographics and AI Usage Experience (including respondents' age, education level, occupation, and duration of use); and (3) Main Constructs, covering MR, PEOU, PU, ATT, and CUI. To ensure data quality, minimum completion times (less than three minutes' duration was considered invalid) and consistency check items were employed to filter out random responses. A total of 419 valid responses were collected.

Descriptive statistics indicated that the majority of respondents were aged between 26 and 41 years, with educational backgrounds primarily in university or technical college. Most respondents

worked in the service industry. In terms of MGAI usage experience, 44% of participants had used MGAI for less than six months, while 40% had used it for six months to one year; this indicates that the sample possesses representativeness and inferential value.

Table 1. Demographic Characteristics

Characteristic	Category	N=419	%
Gender	Male	174	41.5
	Female	245	58.5
Age	Under 25	32	7.6
	26 – 41	282	67.3
	42 – 57	95	22.7
	Above 58	10	2.4
Education Level	Below Junior High School	1	0.2
	High School/Vocational School	39	9.3
	University/College	330	78.8
	Graduate School or Above	49	11.7
Current Occupation	Student	18	4.3
	Information Technology	125	29.8
	Military, Civil Servant, Teacher	38	9.1
	Service Industry	176	42.0
	Finance Industry	29	6.9
	Others	33	7.9
Annual Income (NTD)	0 – 500,000	115	27.4
	500,001 – 1,000,000	232	55.4
	1,000,001 – 1,500,000	54	12.9
	1,500,001 – 2,000,000	12	2.9
	Above 2,010,000	6	1.4
Duration of MGAI Usage	Less than 6 months	167	39.9
	6 months – 1 year	182	43.4
	1 – 2 years	73	16.7

Source: By authors.

4.2 Structural Model Analysis

This study applied SEM to analyze the causal relationships and path effects among the constructs. The results are presented in Table 2. The model fit indices indicated that the chi-square to degrees of freedom ratio (χ^2/df) was 1.664, which is below the recommended threshold of 3, suggesting no significant model misspecification [37]. The root mean square error of approximation (RMSEA) was 0.038, which aligns with the optimal range (< 0.05) as proposed by Browne and Cudeck [38]. The standardized root mean square residual (SRMR) was 0.042, well within the recommended threshold of 0.08 [39]. Additionally, the comparative fit index (CFI) was 0.997 and the Tucker–Lewis index (TLI) was 0.996, both exceeding the recommended benchmark of 0.95 [39]. Overall, the results indicate that the structural model demonstrates a good fit and effectively explains the acceptance mechanisms of MGAI usage behaviors.

Table 2.Model Fit Indices

Fit Index	Value
χ^2	213
χ^2/df	1.664
RMSEA	0.038
SRMR	0.042
CFI	0.997
TLI	0.996

Source: By authors.

4.3 Reliability and Validity Analysis

This study conducted reliability and validity analyses to ensure the robustness of the measurement instruments; the results are presented in Table 3. The internal consistency reliability test showed that, except for ATT, which had a Cronbach's α of 0.693—slightly below the conventional threshold of 0.7—all other constructs (MR, PEOU, PU, and CUI) had Cronbach's α values exceeding 0.7, indicating good internal consistency [38]. The analysis of composite reliability (CR) revealed that all constructs had CR values above 0.7; this confirms that the measurement instruments are stable and effectively capture the intended variables. Moreover, the average variance extracted (AVE) for each construct was greater than 0.5, thus satisfying the standard for convergent validity [40]. Overall, the measurement instruments in this study demonstrate strong reliability and validity, which confirms that they effectively reflect the characteristics of each construct.

Table 3. Reliability and Validity Analysis

Construct	Cronbach's Alpha	CR	AVE
Media Richness (MR)	0.734	0.819	0.532
Attitude (ATT)	0.706	0.815	0.596
Continuous Usage Intention (CUI)	0.734	0.810	0.588
Perceived Ease of Use (PEOU)	0.788	0.849	0.584
Perceived Usefulness (PU)	0.788	0.854	0.594

Source: By authors.

Additionally, this study conducted a heterotrait–monotrait (HTMT) ratio analysis to examine the discriminant validity among the constructs. An HTMT ratio of less than 0.85 indicates good discriminant validity between constructs [41]. The analysis results showed that all HTMT values for the constructs in this study were below the 0.85 threshold, thus meeting the evaluation criteria. This demonstrates that the questionnaire scales exhibit good discriminant validity, to ensure clear differentiation among the constructs. The detailed results are presented in Table 4.

Table 4. HTMT Ratios

	MR	PEOU	PU	ATT	CUI
MR	1.000	0.634	0.789	0.699	0.592
PEOU	0.634	1.000	0.697	0.653	0.510

PU	0.789	0.697	1.000	0.790	0.737
ATT	0.699	0.653	0.790	1.000	0.717
CUI	0.592	0.510	0.737	0.717	1.000

Source: By authors.

4.4 Hypothesis Testing Results and Discussion

This study conducted path analysis using SEM to validate the applicability of the integrated model of TAM and MRT in the context of MGAI acceptance behaviors. The results are presented in Table 5.

The analysis revealed that MR had a significant positive effect on PU ($\beta = 0.45$, $p < 0.01$) and ATT ($\beta = 0.32$, $p < 0.01$), indicating that the real-time feedback and rich information provided by multimodal designs enhance users' value perception of system functionality and strengthen emotional connections. This result is consistent with previous studies, which support the notion that multimodal feedback can enhance functional value [42].

Moreover, PU significantly influenced PEOU ($\beta = 0.38$, $p < 0.01$) and ATT ($\beta = 0.50$, $p < 0.01$), thus supporting the "utility-first, ease-of-use-second" perspective proposed by Wulansari. This finding suggests that when users recognize the benefits of MGAI, they are more willing to invest in learning, which reduces the perception of operational barriers [26]. ATT also significantly predicted CUI ($\beta = 0.62$, $p < 0.01$); this aligns with previous research's conclusions that positive attitudes effectively promote long-term usage tendencies [43].

However, the effect of MR on PEOU was not statistically significant ($\beta = 0.12$, $p > 0.05$). This result may be attributed to the high degree of automation in MGAI, which reduces reliance on interface operations. This finding is consistent with Soenksen et al., who suggested that high-richness multimodal designs do not necessarily improve PEOU in automated systems [44]. Similarly, the effect of PEOU on ATT was also insignificant ($\beta = 0.09$, $p > 0.05$), indicating that experienced users are more focused on functional value (PU) rather than operational simplicity. This phenomenon aligns with the findings of Heirati, who noted that technical familiarity shifts the evaluation model towards a focus on technological benefits [19]. Wang further pointed out that as technological familiarity increases, the impact of operational complexity on users gradually diminishes [22].

Descriptive statistics showed that 83.3% of participants had used MGAI for more than six months (Table 1); this suggests that technological familiarity may weaken the association between MR and PEOU, while highlighting the critical role of PU and ATT in long-term usage. This result implies that as user experience accumulates, there is a greater emphasis on system performance and practical application rather than operational simplicity.

Table 5. Path Analysis Results

Hypothesis	Path	Estimate	Standard Error (SE)	Standardized Coefficient (β)	z-value	p-value	Result
H1	MR->PU	0.9482	0.0506	0.8484	18.756	< .001	Supported
H2	MR->ATT	0.2253	0.0982	0.1957	2.293	0.022	Supported

H3	MR->PEOU	0.0863	0.1102	0.0819	0.783	0.434	Not Supported
H4	PU->PEOU	0.6547	0.1012	0.6950	6.472	< .001	Supported
H5	PEOU->ATT	0.0387	0.0760	0.0354	0.509	0.611	Not Supported
H6	PU->ATT	0.7498	0.1152	0.7279	6.510	< .001	Supported
H7	ATT->CUI	0.9906	0.0353	1.0116	28.025	< .001	Supported

Source: By authors.

5. Conclusion and Recommendations

5.1 Conclusion

This study successfully extends the TAM and MRT to analyze user acceptance behavior toward MGAI; it also proposes an integrated theoretical framework to explain user behavior in relation to multimodal technologies. The empirical results indicate that MR significantly enhances PU; this supports the theoretical assertions of Daft and Lengel [10], which suggest that richer media can improve the effectiveness of information transmission and strengthen users' perceptions of system value. In multimodal scenarios, MGAI's integration of voice, images, and text interactions significantly enhances users' recognition of system efficiency, thereby reinforcing their intention for long-term usage. Compared to traditional unimodal technologies, MGAI exhibits greater interactivity and richer information transmission; it enables users to better perceive the technology's value, which in turn positively influences their decision to continue using the system.

Although MR effectively enhances PU, the results of this study reveal that MR does not significantly impact PEOU. This finding contradicts the traditional perspective of TAM, which assumes that a more intuitive and media-rich system should enhance user operation and learning efficiency. However, empirical evidence suggests that in highly multimodal MGAI environments, the simultaneous presentation of rich media does not significantly reduce users' cognitive load. In fact, the parallel operation of multiple media channels may increase cognitive demands [19]. This implies that in the context of MGAI applications, MR is not the key factor in simplifying operation; instead, it may elevate users' comprehension costs. When information is derived from various media and requires simultaneous processing, users are more likely to experience cognitive overload.

Additionally, this study finds that PEOU does not significantly affect ATT. According to TAM's theoretical assumptions, PEOU should have a significant impact on user attitudes, as simple and intuitive operations generally enhance user acceptance. However, this correlation was not validated in the multimodal MGAI environment [22]. A possible explanation is that users engaging with technology in a media-rich multimodal setting are more focused on the functional benefits and real-time feedback provided by the technology, rather than the simplicity of its operation. When MGAI offers highly integrated and diverse information, users tend to overlook operational complexity because they perceive that the richness of information far outweighs the challenges of usability. Thus, in the application of MGAI, users' attitudes are primarily driven by the practical effectiveness

provided by the technology, rather than merely the ease of operation.

These findings challenge the assumption in TAM that ease of use is a primary acceptance indicator for unimodal technologies; this demonstrates that in the design of MGAI, users prioritize the functional value and media richness of the technology over simplistic interface designs. Thus, designers should reconsider interface design in multimodal contexts, to emphasize the synchronous collaboration of rich media and the accuracy of information transmission. At the same time, attention should be paid to reducing excessive cognitive load in multimodal integration scenarios, to improve user comprehension and acceptance.

The empirical results of this study not only extend the explanatory power of TAM and MRT in the field of multimodal technologies, but also propose new directions for multimodal design. This emphasizes the need to balance interaction richness with information clarity, to ensure that increased media interactions do not lead to excessive information interference, and thus ultimately achieve an optimal user experience.

5.2 Managerial Implications

The findings of this study provide several insights into the design and management strategies of MGAI.

First, the results indicate that MR significantly influences PU. This implies that in the development process of MGAI, designers should prioritize the enhancement of multimodal media integration, including the seamless synchronization of voice, images, and text. Through rich media interaction designs, users can receive, process, and comprehend information more efficiently, thereby enhancing their sense of recognition and trust in the technology [24]. This is particularly evident in educational and medical contexts, where real-time multimodal feedback can substantially improve interactive efficiency and the accuracy of information transmission. For example, in intelligent education systems, the integration of voice explanations, visualized learning materials, and real-time interactions can effectively promote learning outcomes. Similarly, in telemedicine, the synchronous integration of voice diagnostics and visual aids can assist medical professionals in performing faster and more accurate diagnoses.

However, the study also reveals that MR does not significantly enhance PEOU. This presents a critical challenge for managers and design teams: how to increase multimodal richness while simultaneously reducing users' learning barriers and operational complexity. Designers should consider simplifying user interfaces, by avoiding excessive operational steps and redundant interaction processes. Additionally, contextual guidance and intelligent assistance can be employed to help users adapt to the operating environment more quickly [19].

Furthermore, the findings show that PEOU does not have a significant impact on ATT; this suggests that when users decide whether to adopt MGAI, they prioritize the functional benefits of the technology over its operational simplicity. Therefore, technology developers should emphasize the core functional advantages of MGAI in their market strategies—including data processing capabilities, real-time feedback, and the efficiency of multimodal collaboration—rather than focusing

solely on interface simplicity. This strategic focus is more likely to align with target market demands, and will thereby enhance technology acceptance and long-term usage intentions.

5.3 Practical Implications

The findings of this study provide concrete practical recommendations for the application of MGAI across different industries.

First, in the education sector, multimodal technologies can significantly enhance the richness of teaching interactions and improve learning efficiency. Based on the study's findings, which indicate that MR positively influences PU, educational institutions could consider adopting intelligent learning platforms that integrate voice assistance, image recognition, and interactive learning materials to enable real-time interaction and personalized learning [20]. Such multimodal learning systems can better accommodate the diverse needs of different learners and enhance learning outcomes through multi-sensory engagement. Additionally, mixed reality technology has been successfully applied in nuclear energy education, using 3D interactive content to deepen students' understanding and increase engagement, demonstrating the potential of multimodal approaches in science education[45].

In the healthcare sector, MGAI's multimodal integration technology can be applied to telemedicine, clinical diagnostics, and health monitoring. The findings of this study indicate that media richness (MR) enhances the efficiency and accuracy of information transmission, suggesting that synchronized voice descriptions, image recognition, and data analysis can enable medical professionals to perform more accurate diagnoses and provide real-time feedback [19]. For instance, a recent study developed a deep learning–based burn injury classification platform that integrates image recognition with a web-based interface. Users can upload wound images via mobile devices, and the system automatically identifies the severity level and provides relevant medical suggestions. The platform was specifically designed for non-expert users, emphasizing real-time feedback and operational simplicity. It demonstrated high classification accuracy and strong user satisfaction [46]. Furthermore, another study introduced the "Beauty Model—Professional Skin Condition AI Detection Platform," which similarly combines web-based multimodal AI with user-centric interaction. This platform enables users to upload facial skin images and receive automatic diagnostic feedback on conditions such as eczema or dermatitis. Designed to reduce appearance-related anxiety, it provides low-barrier access, real-time feedback, and clear guidance—all of which improve perceived usefulness and trust in AI recommendations[47]. These cases exemplify how multimodal AI applications in medical contexts not only enhance perceived usefulness and user experience but also reinforce continued usage intention—echoing this study's empirical findings on the positive relationship between media richness and perceived usefulness.

In the commercial domain, MGAI can enhance customer service and consumer experiences through rich multimodal interactions. The study results demonstrate that MR effectively increases user perceptions of technology's usefulness. Therefore, on e-commerce platforms, intelligent customer service systems that integrate voice navigation, image-based recommendations, and text-

based interactions can significantly improve the shopping experience [24]. Such designs not only increase customer satisfaction, but also enhance shopping fluidity and conversion rates.

Overall, MGAI's multimodal design showcases greater interactive value and efficiency in information transmission across education, healthcare, and business applications. Designers and developers should focus on balancing MR with the user experience, to ensure both acceptance and long-term development of the technology in practical applications.

5.4 Future Research

Despite the successful validation of the user acceptance model for MGAI and the significant association identified between MR and PU, this study has several limitations that warrant consideration.

First, the data collected for this study are based on a cross-sectional survey at a single point in time, which makes it difficult to observe how user acceptance behaviors toward MGAI evolve over time. Future research could consider conducting longitudinal studies to explore how users' PU and PEOU may change with accumulated experience during extended interactions with MGAI. Such studies would offer deeper insights into the temporal dynamics of user acceptance.

Second, the sample of this study predominantly consists of a specific user group, and lacks broad representation across different age groups, educational backgrounds, and levels of technological familiarity. This limitation may affect the generalizability of the findings. Future studies should aim to broaden the diversity of samples, particularly by including older adults and users without a technological background, to achieve a more comprehensive understanding of MGAI acceptance behaviors.

Additionally, the study did not find significant evidence to support the impact of MR on PEOU, which may reflect the cognitive load introduced by multimodal integration. Future research should further investigate how to reduce cognitive load and optimize the fluidity of multimedia interactions, especially concerning highly complex technological operations. Moreover, examining cross-cultural differences in user acceptance behaviors could provide valuable insights into how MGAI is perceived in global markets. Cross-cultural comparative studies would enable a better understanding of MGAI's acceptance across different cultural contexts.

Finally, given the rapid development of the internet of things and augmented reality technologies, future research could explore the integration of MGAI with these technologies, to understand their collective impact on user acceptance behaviors. Such integration is particularly promising in the contexts of smart cities, smart healthcare, and interactive learning. The convergence of these technologies could further enhance the practical applications of multimodal technologies and provide richer interactive experiences.

References

[1] Joshi, G., Walambe, R. and Kotecha, K. A Review on Explainability in Multimodal Deep Neural Nets. IEEE Access, 2021. DOI: 10.1109/access.2021.3070212.

[2] Zong, Y. Multimodal Imaging in Oncology: Challenges and Future Directions. Academic Journal of Science and

Technology, 2024. DOI: 10.54097/qw4tja89.

[3] Niu, W. The Role of Artificial Intelligence Autonomy in Higher Education: A Uses and Gratification Perspective. Sustainability, 2024. DOI: 10.3390/su16031276.

[4] Chin, C.-H. Exploring the Usage Intention of AI-powered Devices in Smart Homes Among Millennials and Zillennials: The Moderating Role of Trust. Young Consumers Insight and Ideas for Responsible Marketers, 2023. DOI: 10.1108/yc-05-2023-1752.

[5] Tortora, L. Beyond Discrimination: Generative AI Applications and Ethical Challenges in Forensic Psychiatry. Frontiers in Psychiatry, 2024. DOI: 10.3389/fpsyt.2024.1346059.

[6] Yang, J. Development and Challenges of Generative Artificial Intelligence in Education and Art. Highlights in Science Engineering and Technology, 2024. DOI: 10.54097/vaeav407.

[7] Lee, J.C. and Lin, R. The Continuous Usage of Artificial Intelligence (AI)-powered Mobile fitness Applications: The goal-Setting Theory Perspective. Industrial Management & Data Systems, 2023. DOI: 10.1108/imds-10-2022-0602.

[8] Davis, F.D. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. MIS Quarterly, 1989, 13(3), 319–340. DOI: 10.2307/249008.

[9] Tahar, A., Riyadh, H.A., Sofyani, H. and Purnomo, W.E. Perceived Ease of Use, Perceived Usefulness, Perceived Security and Intention to Use E-Filing: The Role of Technology Readiness. Journal of Asian Finance Economics and Business, 2020. DOI: 10.13106/jafeb.2020.vol7.no9.537.

[10] Daft, R.L. and Lengel, R.H. Organizational Information Requirements, Media Richness and Structural Design. Management Science, 1986, 32(5), 554–571. DOI: 10.1287/mnsc.32.5.554.

[11] Zhang, M., Lin, W.S., Zhen, M., Yang, J. and Zhang, Y. Users' Health Information Sharing Intention in Strong Ties Social Media: Context of Emerging Markets. Library Hi Tech, 2021. DOI: 10.1108/lht-02-2020-0024.

[12] Radford, A., et al. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning, PMLR, 2021, 8748–8763. DOI: (accessed May 10, 2025), http://proceedings.mlr.press/v139/radford21a.

[13] Wu, S., Fei, H., Qu, L., Ji, W. and Chua, T.-S. Next-gpt: Any-to-any multimodal LLM. In Forty-first International Conference on Machine Learning, 2024. Accessed May 10, 2025. https://openreview.net/forum?id=NZQkumsNlf.

[14] Joshi, M., Chen, D., Liu, Y., Weld, D.S., Zettlemoyer, L. and Levy, O. SpanBERT: Improving pre-training by representing and predicting spans. Transactions of the Association for Computational Linguistics, 2020, 8, 64–77.

[15] Liu, X., Wang, X., Li, J. and Chen, M. The Effect of Media Richness on the Stability of Physician-Patient Relationships on E-Consultation Platforms. Journal of Global Information Management, 2022. DOI: 10.4018/jgim.315301.

[16] Alayrac, J.-B., et al. Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems, 2022, 35, 23716–23736.

[17] Chen, P., Zhang, Z., Dong, Y., Zhou, L. and Wang, H. Enhancing Visual Question Answering through Ranking-Based Hybrid Training and Multimodal Fusion. Journal of Intelligence Technology and Innovation, 2024, 2(3), Oct. DOI: 10.30212/JITI.202402.011.

[18] Ho, J., Jain, A. and Abbeel, P. Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 2020, 33, 6840–6851.

[19] Heirati, N. When the recipe is more important than the ingredients: Unveiling the complexity of consumer use of voice assistants. Psychology and Marketing, 2024. DOI: 10.1002/mar.21992.

[20] Yao, L. and Liu, Y. Emotional multifaceted feedback on AI tool use in EFL learning initiation: Chain-mediated effects of motivation and metacognitive strategies in an optimized TAM model. arXiv, 2025. DOI: 10.48550/arXiv.2503.18180.

[21] Hu, J.J., Liu, M.H., Wen, J.R. and Doong, S.H. Investigating the impact of peer relationships on student's motivation and attitudes towards mobile gaming. Journal of Information and Computing, 2025, 2(4). DOI: 10.30211/JIC.202402.009.

[22] Wang, Z. Media richness and continuance intention to online learning platforms: The mediating role of social presence and the moderating role of need for cognition. Frontiers in Psychology, 2022, 13, 950501.

[23] Venkatesh, V. and Bala, H. Technology acceptance model 3 and a research agenda on interventions. Decision Sciences, 2008, 39(2), 273–315. DOI: 10.1111/j.1540-5915.2008.00192.x.

[24] Chatterjee, S., Chaudhuri, R., Vrontis, D., Thrassou, A. and Ghosh, S.K. Adoption of artificial intelligence-integrated CRM systems in agile organizations in India. Technological Forecasting and Social Change, 2021, 168, 120783.

[25] Moon, J., Myungkeun, S., Lee, W.S. and Shim, J.M. Structural relationship between food quality, usefulness, ease of use, convenience, brand trust and willingness to pay: The case of Starbucks. British Food Journal, 2022. DOI: 10.1108/bfj-07-2021-0772.

[26] Wulansari, K. The influence of digital literacy on intention to use QRIS by using TAM as the cashless paying method on MSME in Samarinda Seberang District. Kne Social Sciences, 2024. DOI: 10.18502/kss.v9i11.15840.

[27] Karki, S. and Dahal, A.R. Awareness and adoption of the cashless economy in Nepal. International Journal of Management and Organization, 2024, 2(2). DOI: 10.30209/IJMO.202402.002.

[28] Lu, K., Pang, F. and Shadiev, R. Understanding college students' continuous usage intention of asynchronous online courses through extended technology acceptance model. Education and Information Technologies, 2023. DOI: 10.1007/s10639-023-11591-1.

[29] Harnida, M. and Mardah, S. Penerapan technology acceptance model terhadap perilaku pengguna uang elektronik. Al-Kalam Jurnal Komunikasi Bisnis Dan Manajemen, 2023. DOI: 10.31602/al-kalam.v10i1.9019.

[30] Chung, J.B. and Nam, S.J. A study on the user acceptance of O2O services: Mediating effect of customer attitude. East Asia Business and Economics Association, 2020. DOI: 10.20498/eajbe.2020.8.3.15.

[31] Yuan, D., Rahman, M.K., Gazi, M.A.I., Rahaman, M.A., Hossain, M.M. and Akter, S. Analyzing of user attitudes toward intention to use social media for learning. Sage Open, 2021. DOI: 10.1177/21582440211060784.

[32] Ajzen, I. The theory of planned behavior. Organizational Behavior and Human Decision Processes, 1991, 50(2), 179–211.

[33] Yang, Z., Zhou, Q., Chiu, D.K.W. and Wang, Y. Exploring the factors influencing continuous usage intention of academic social network sites. Online Information Review, 2022. DOI: 10.1108/oir-01-2021-0015.

[34] Tseng, C.H. and Wei, L.F. The efficiency of mobile media richness across different stages of online consumer behavior. International Journal of Information Management, 2020, 50, 353–364.

[35] Venkatesh, V. and Davis, F.D. A theoretical extension of the Technology Acceptance Model: Four longitudinal field studies. Management Science, 2000. DOI: 10.1287/mnsc.46.2.186.11926.

[36] Chen, Q., Chen, M., Wei, Z., Wang, G., Ma, X. and Evans, R. Unpacking the black box: How to promote citizen

engagement through government social media during the COVID 19 crisis. Computers in Human Behavior, 2020. DOI: 10.1016/j.chb.2020.106380.

[37] Kline, R.B. Principles and practice of structural equation modeling. Guilford Publications, 2023. Accessed May 11, 2025. https://books.google.com/books?hl=zhTW&lr=&id=t2CvEAAAQBAJ&oi=fnd&pg=PP1&dq=Principles+and+Practice+of+Structural+Equation+Modeling+(4th+ed.).+New+York,+NY:+Guilford+Press.&ots=sWVF-7c1eK&sig=zIHlvgf6ejnClRIAmO786Oc5KyQ

[38] Browne, M.W. and Cudeck, R. Alternative ways of assessing model fit. Sociological Methods & Research, 1992, 21(2), 230–258. DOI: 10.1177/0049124192021002005.

[39] Hu, L. and Bentler, P.M. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal, 1999, 6(1), 1–55. DOI: 10.1080/10705519909540118.

[40] Fornell, C. and Larcker, D.F. Evaluating structural equation models with unobservable variables and measurement error. Journal of Marketing Research, 1981, 18(1), 39–50. DOI: 10.1177/002224378101800104.

[41] Henseler, J., Ringle, C.M. and Sarstedt, M. A new criterion for assessing discriminant validity in variance based structural equation modeling. Journal of the Academy of Marketing Science, 2015, 43(1), 115–135. DOI: 10.1007/s11747-014-0403-8.

[42] Mohsen, F., Ali, H., Hajj, N.E. and Shah, Z. Artificial intelligence based methods for fusion of electronic health records and imaging data. Scientific Reports, 2022. DOI: 10.1038/s41598-022-22514-4.

[43] Rahmiati, R., Rozi, I.F., Septrizola, W., Sarianti, R. and Patrisia, D. Determinants of actual digital library usage. 2019. DOI: 10.2991/icebef-18.2019.166.

[44] Soenksen, L.R. and others. Integrated multimodal artificial intelligence framework for healthcare applications. NPJ Digital Medicine, 2022. DOI: 10.1038/s41746-022-00689-4.

[45] Huang, Y.-P. Combined the mixed reality technology in design an interactive learning environment. International Journal of Management and Organization, 2024, 2(3). DOI: 10.30209/IJMO.202402.008.

[46] Sun, P.-L., et al. Real time AI image classification system for burn injuries. Journal of Information and Computing, 2024, 2. DOI: 10.30211/JIC.202402.001.

[47] Lin, S.-H., et al. Enhancing skin health in post-epidemic with Beauty Model Professional Skin Condition AI Detection Platform. Journal of Information and Computing, 2024, 2. DOI: 10.30211/JIC.202402.002.