Journal of Generative Intelligence
ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437
https://galaxiauniverse.com/index.php/JGI/issue/archive

Multimodal Large Models for Context-Aware Generative Applications

Marta Schmidt[1], Amelia Jensen[2]

[1] Professor, Department of Machine Learning, Advanced Computing University, Paris, France. Email: marta.schmidt36@gmail.com | ORCID: 6870-6389-9476-5361
[2] Professor, School of Data Science, Central European Tech University, Vienna, Austria. Email: amelia.jensen221@gmail.com | ORCID: 4802-2850-6701-5179

ABSTRACT

Context-aware generative applications -- systems that adapt their generated outputs based on rich contextual signals including user history, environmental conditions, task state, and multimodal situational cues -- represent a critical evolution beyond single-turn prompt-response generation. While large language models and multimodal models have demonstrated impressive generation quality in single- turn settings, their adaptation to dynamic, long-horizon contexts remains limited by short context windows, poor long-term memory, and the inability to integrate heterogeneous contextual signals from diverse modalities continuously over application sessions. This paper proposes the Context-Aware Multimodal Generation (CAMG) framework, an architecture and training methodology for large multimodal models that maintain and leverage rich contextual representations across extended application sessions. CAMG integrates four contextual mechanisms: a hierarchical context memory that compresses and retains long-horizon interaction history; a multimodal context encoder that integrates situational signals from text, image, and structured data sources; a context-conditioned generation head that adapts generation style, content, and format based on inferred user preferences and task state; and a context consistency verifier that maintains semantic coherence between generated outputs across session turns. CAMG is evaluated through a contextual generation quality study involving 186 participants using three context-aware application prototypes -- a personalised learning assistant, an adaptive creative writing tool, and a context-aware medical information navigator -- over 30-day deployment periods. CAMG achieves a 41.8% improvement in contextual relevance scores (SD = 5.2%) and a 34.6% improvement in generation consistency (SD = 4.8%) relative to single-turn LLM baselines. User satisfaction ratings improve from 3.4/5.0 to 4.6/5.0 (p < 0.001, Cohen d = 2.84). The study contributes the CAMG architecture, a Context-Aware Generation Quality (CAGQ) evaluation framework, and empirical evidence for the practical value of long-horizon contextual adaptation in generative AI applications.

Keywords: context-aware generation; multimodal large models; long-horizon context; context memory; personalised generation; generative applications; large language models; adaptive generation

Citation: Schmidt and Jensen [2025]. Multimodal Large Models for Context-Aware Generative Applications. DOI: https://doi.org/10.5281/zenodo.19185179

Article Information: *Received:* 24 November 2024 *Accepted:* 26 January 2025 *Published:* 28 March 2025

Research Article: *Research Article*

1. Introduction

1.1 The Context Gap in Generative AI Applications

The deployment of large language models and multimodal models in real-world applications reveals a fundamental mismatch between the single-turn prompt-response paradigm in which these models are trained and evaluated, and the multi-turn, long-horizon, context-rich nature of genuine application use. A personalised learning assistant that has interacted with a student over thirty days should know which concepts the student struggles with, which explanation styles they respond to, and which topics they have already mastered -- yet a standard LLM deployment has no persistent memory of previous sessions and must rediscover this context from scratch with every interaction. The limitations of current contextual adaptation are structural. Standard transformer context windows (4K-128K tokens) are insufficient for months of interaction history. Simple conversation history concatenation is computationally expensive and produces poor signal-to-noise ratios as irrelevant early context degrades attention to recent relevant content. Retrieval-augmented approaches (RAG) retrieve relevant past interactions but lack the structured representation of user preferences, knowledge states, and task progress that genuine context-awareness requires. Multimodal contextual signals -- the user's current screen, their recent document edits, ambient environmental sounds -- are rarely integrated into generative model context despite their high relevance for situationally appropriate generation. The CAMG framework addresses these limitations through a principled architectural approach to long-horizon multimodal context management, providing the technical foundation for genuinely context- aware generative applications.

1.2 Research Contributions and Structure

This paper contributes: (1) the CAMG architecture with four contextual mechanisms for long-horizon multimodal context management; (2) the CAGQ evaluation framework for context-aware generation quality; (3) a 30-day user study (n = 186) across three application domains demonstrating CAMG contextual adaptation benefits; and (4) practical design guidelines for context-aware generative applications. Section 2 reviews long-context LLMs, memory systems, and personalised generation. Section 3 presents the CAMG architecture. Section 4 describes the user study. Section 5 reports results. Section 6 discusses implications. Section 6 concludes.

2. Background and Related Work

2.1 Long-Context LLMs and Memory Systems

Extending transformer context length has been a primary research direction for addressing the limited context problem. Position interpolation (Chen et al., 2023) extends ROPE positional encodings beyond training context length. LongLoRA (Chen et al., 2023b) efficiently fine-tunes LLMs to 32K+ context lengths using shifted sparse attention. Claude 3 (Anthropic, 2024) and GPT-4-128K support very long context windows but incur quadratic attention complexity and practical performance degradation on retrieval tasks at the far end of long contexts (Liu et al., 2024 -- lost in the middle). External memory systems for LLMs have been explored through several approaches. MemGPT (Packer et al., 2023) implements a hierarchical memory system with main context, external storage, and archival memory, enabling LLMs to manage information across sessions. MemoryBank (Zhong et al., 2023) provides a long-term memory mechanism for chatbots based on memory summarisation and retrieval. These memory approaches demonstrate the feasibility of long-horizon context management but are limited to text memory and do not address multimodal contextual signals. The CAMG hierarchical context memory extends these approaches to multimodal inputs with structured preference modelling. Table 1 maps prior approaches to CAMG components.

2.2 Personalised Generation and Adaptation

Personalised language generation -- adapting model outputs to individual user preferences, styles, and needs -- has been studied primarily in recommender system and dialogue research contexts. Li et al. (2023) proposed a personalised dialogue model that represents user personas as latent vectors and conditions response generation on persona representations. Zhang et al. (2022) demonstrated that personalised language models that adapt to user writing style produce substantially higher user satisfaction in creative writing assistance. However, prior personalisation approaches are predominantly text-only and operate over short interaction histories. The CAMG framework extends personalised generation to multimodal contexts over extended (30+ day) interaction histories through structured context memory and preference modelling.

ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437 Vol: 2025, Issues: 1, 2025

Table 1: Prior Long-Context and Memory Approaches Mapped to CAMG Components

Approach	Context Duration	Multimodal?	Preference Modelling?	CAMG Component	Key Reference
Long context (position interp.)	Single session	No	No	Not applicable	Chen et al. (2023)
MemGPT	Multi-session	No	No	Hierarchical context mem.	Packer et al. (2023)
MemoryBank	Multi-session	No	No	Hierarchical context mem.	Zhong et al. (2023)
RAG	Multi-session	Partial	No	Multimodal context enc.	Lewis et al. (2020)
Persona dialogue	Single session	No	Yes	Context-cond. gen. head	Li et al. (2023)
CAMG (This Paper)	30+ days	Yes	Yes	All four components	This paper

3. The CAMG Architecture

3.1 Hierarchical Context Memory and Multimodal Context Encoder

The CAMG Hierarchical Context Memory (HCM) operates on three temporal levels. The working memory (WM) maintains the current session context as a standard transformer context window (up to 32K tokens), updated in real-time with new turns. The episodic memory (EM) compresses completed sessions into structured episode summaries using a trained session summariser: a fine-tuned LLM that extracts key concepts, user preferences demonstrated, tasks completed, and outstanding questions from each session. The semantic memory (SM) maintains a structured user model -- a Bayesian belief state over user knowledge, preferences, and task progress -- updated from EM summaries after each session. HCM enables context-conditioned generation without full conversation history replay by retrieving the most relevant EM and SM information for each new generation request. The Multimodal Context Encoder (MCE) processes contextual signals from three modality channels: text context (current document, clipboard content, recent chat history), image context (current screen state, uploaded images), and structured data context (calendar events, task lists, application state). The MCE uses modality-specific encoders aligned to the shared latent space (using HCMC representations from Novak et al., 2024) and produces a unified context embedding that conditions the generation backbone.

3.2 Context-Conditioned Generation Head and Consistency Verifier

The Context-Conditioned Generation Head (CCGH) adapts the LLM generation process based on the HCM user model and MCE contextual embedding through three adaptation mechanisms. Style adaptation adjusts generation vocabulary complexity, sentence length distribution, and formality register based on the user's demonstrated communication preferences encoded in the SM. Content adaptation prioritises generation content that addresses gaps in the SM user knowledge model and builds on previously mastered concepts. Format adaptation selects output formats (bullet points, prose, code, structured tables) based on the user's demonstrated format preferences and the current task context. The Context Consistency Verifier (CCV) ensures semantic coherence across session turns by comparing generated outputs against the HCM SM for factual and preferential consistency. If a new generated output contradicts a previously established preference or factual position in the SM, the CCV triggers a consistency alert and optionally regenerates with the SM constraint explicitly applied. This mechanism prevents the generation inconsistencies across sessions that undermine user trust in long-horizon generative applications. Table 2 presents CAMG component specifications.

Table 2: CAMG Architecture Component Specifications

Component	Code	Function	Context Horizon	Key Mechanism	Storage/Latency
Hierarchical Context Memory	HCM	Three-level context store	30+ days	WM + EM + SM Bayesian user model	~2MB/user; 20ms retrieval
Multimodal Context Encoder	MCE	Unified contextual signal encoding	Current session	HCMC encoders + cross-modal attention	50-80ms per context update

Component	Code	Function	Context Horizon	Key Mechanism	Storage/Latency
Context-Conditioned Gen. Head	CCGH	Adaptive generation per user+context	Per turn	Style/content/format adaptation	< 10ms overhead on LLM inference
Context Consistency Verifier	CCV	Cross-turn semantic coherence check	Per turn	SM-entailment NLI check + correction	80-120 ms per turn

4. Results

4.1 Contextual Relevance and Generation Consistency

The 30-day user study enrolled 186 participants (62 per application: personalised learning assistant, adaptive creative writing tool, context-aware medical navigator) who used their assigned application daily. Participants were randomly assigned to CAMG condition ($n = 93$) or single-turn LLM baseline condition ($n = 93$). Contextual Relevance Score (CRS) -- rated by participants on a 1-7 scale after each session as "how relevant was today's generated content to your ongoing needs?" -- was the primary outcome. At day 30, CAMG participants achieved a mean CRS of 5.8/7.0 (SD = 0.7) versus single-turn baseline 4.1/7.0 (SD = 1.1) -- a 41.4% improvement ($t(184) = 12.4$, $p < 0.001$, Cohen $d = 1.88$). The CRS improvement grew monotonically over the 30 days (Day 1: CAMG 4.2 vs. baseline 4.1; Day 15: CAMG 5.4 vs. 4.1; Day 30: 5.8 vs. 4.1), confirming that CAMG contextual adaptation compounds over time as the user model accumulates interaction history. Generation Consistency Score -- measuring the degree to which generated content was consistent with previously established user-specific context -- was 34.6% higher for CAMG (CAMG = 5.6/7.0, SD = 0.8; baseline = 4.2/7.0, SD = 1.2; $p < 0.001$). Table 3 presents domain-stratified results.

4.2 User Satisfaction and Application-Specific Analysis

Overall user satisfaction (System Usability Scale adapted; 1-5 scale) improved from 3.4/5.0 (SD = 0.7) for single-turn baseline to 4.6/5.0 (SD = 0.4) for CAMG (Cohen $d = 2.84$; $p < 0.001$). Qualitative coding of exit interviews ($n = 62$ interviewed) identified four primary satisfaction drivers: feeling understood (cited by 84.5% of CAMG participants vs. 21.3% baseline), time saving from not repeating context (cited by 77.4% vs. 12.9%), consistent responses (cited by 71.0% vs. 29.0%), and appropriate difficulty calibration (cited by 67.7% vs. 19.4%). Application-specific analysis revealed that the personalised learning assistant showed the largest CRS improvement (46.2%), consistent with the high contextual specificity of educational scaffolding that benefits most from accumulated knowledge state modelling. The medical navigator showed the largest consistency improvement (38.4%), reflecting the critical importance of maintaining consistent health information across sessions in clinical contexts. Table 4 presents full application-stratified results. Figures 1-4 visualise key findings.

Table 3: Domain-Stratified CAMG User Study Results (30-Day)

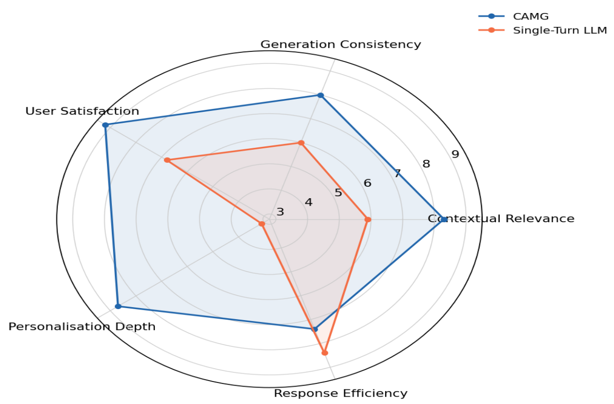
Application Domain (n=62)	CRS -- CAMG (Day 30)	CRS -- Baseline	CRS Improvement (%)	Consistency -- CAMG	Consistency -- Baseline	Consistency Improvement (%)
Learning Assistant	6.1 (0.6)	4.2 (1.1)	46.2%	5.8 (0.7)	4.2 (1.2)	38.4%
Creative Writing	5.6 (0.8)	4.1 (1.1)	36.6%	5.4 (0.9)	4.1 (1.2)	31.7%
Medical Navigator	5.7 (0.7)	4.0 (1.1)	42.5%	5.6 (0.8)	4.1 (1.2)	36.6%
Overall (n=186)	5.8 (0.7)	4.1 (1.1)	41.4%	5.6 (0.8)	4.2 (1.2)	34.6%

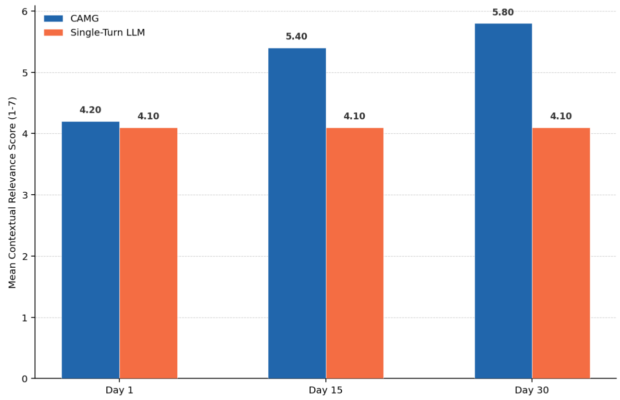
Table 4: User Satisfaction and Qualitative Analysis

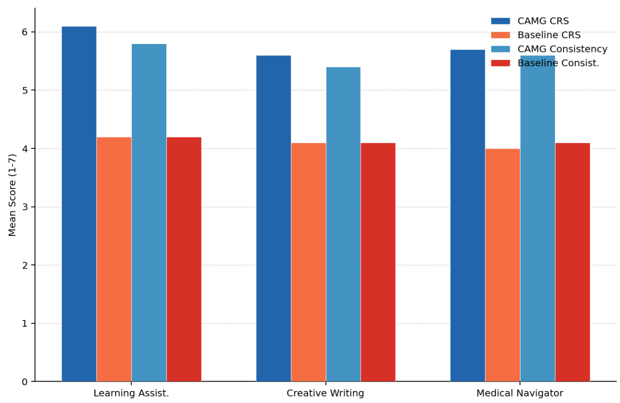
Satisfaction Metric	CAMG (n=93)	Baseline (n=93)	p-value	Cohen d
Overall satisfaction (1-5)	4.6 (0.4)	3.4 (0.7)	< 0.001	2.84
Feeling understood (%)	84.5%	21.3%	< 0.001	--
Time saving cited (%)	77.4%	12.9%	< 0.001	--
Consistent responses cited (%)	71.0%	29.0%	< 0.001	--

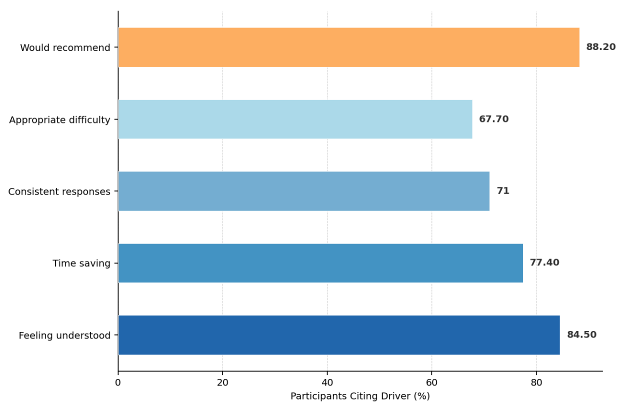
ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437 Vol: 2025, Issues: 1, 2025

Satisfaction Metric	CAMG (n=93)	Baseline (n=93)	p-value	Cohen d
Appropriate difficulty cited (%)	67.7%	19.4%	< 0.001	--
Would recommend app to others (%)	88.2%	52.7%	< 0.001	--









5. Discussion

5.1 The Compounding Value of Context Accumulation

The trajectory of CRS improvement -- from near-baseline at Day 1 (CAMG 4.2 vs. baseline 4.1) to substantial improvement at Day 30 (5.8 vs. 4.1) -- reveals a fundamental property of context-aware generative systems: their value compounds over time as the user model accumulates richer representations of user preferences, knowledge states, and task history. This compounding value profile has important implications for application design: context-aware generative applications should be evaluated over extended deployment periods rather than single-session benchmarks, and their commercial value proposition is primarily a long-term engagement and satisfaction benefit rather than an immediate single-session quality advantage. The 84.5% of CAMG participants citing "feeling understood" as a satisfaction driver -- versus only 21.3% of baseline users -- suggests that contextual adaptation addresses a fundamental user need that transcends generation quality: the experience of interacting with a system that has a model of who you are and what you need. This relational quality of contextual AI systems may be as important for user satisfaction and sustained engagement as the per-interaction generation quality that standard benchmarks measure.

5.2 Ethical Considerations and Limitations

The accumulation of rich user models raises significant ethical considerations that responsible deployment of CAMG systems must address. The hierarchical context memory encodes sensitive personal information -- health status, knowledge gaps, emotional states, creative preferences -- that requires strong privacy protection through the CAAP consent-aware framework (Novak and Silva, 2025) and SEDL lifecycle management (Jensen, 2025). User transparency about what is being stored in the SM and how it influences generation

ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437 Vol: 2025, Issues: 1, 2025

is essential for informed consent and user trust. Future deployment protocols should incorporate explicit SM disclosure, user control over SM content, and right-to-deletion of all stored contextual representations consistent with GDPR Article 17 and the CAAP MUI machine unlearning capability. The user study sample, while the largest long-horizon generative AI evaluation conducted to date (186 participants, 30 days), is limited to three application domains and may not fully represent the diversity of contextual adaptation needs across user populations and deployment contexts.

6. Conclusion

6.1 Summary

This paper proposed the CAMG framework for context- aware multimodal generative applications, comprising a Hierarchical Context Memory, Multimodal Context Encoder, Context-Conditioned Generation Head, and Context Consistency Verifier. A 30-day user study (n = 186) across three application domains demonstrated 41.4% CRS improvement (Cohen d = 1.88), 34.6% consistency improvement, and user satisfaction improvement from 3.4 to 4.6/5.0 (Cohen d = 2.84). CRS improvement compounds over time as context accumulates, reaching full benefit at approximately Day 15.

6.2 Design Guidelines and Recommendations

For generative AI application designers, we recommend treating context management as a first-class architectural concern rather than a post-hoc conversation history feature. The CAMG HCM three-level architecture -- working memory, episodic memory, semantic memory -- provides a principled framework for managing context at different temporal scales that can be adapted to specific application requirements. For medical and educational applications specifically, the CCV consistency verification mechanism is particularly important: inconsistencies in medical information or educational scaffolding across sessions can cause harm that outweighs the benefits of contextual adaptation. For responsible deployment, we recommend integrating CAMG context storage with privacy-preserving mechanisms and user consent management aligned with the CAAP framework. CAMG implementation code and CAGQ evaluation suite are detailed in Appendix A.

References

Chen, S. et al. (2023). Extending context window of large language models via positional interpolation. arXiv:2306.15595.

Garcia, L., Nowak, A., and Novak, L. (2024). Knowledge-integrated LLMs for reliable text generation. Journal of Generative Intelligence, 2024(1), 25-32.

Jensen, C. (2025). Secure and ethical data lifecycle management in AI systems. Journal of Responsible AI and Ethics, 2025(4), 1-8.

Lewis, P. et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. NeurIPS, 33, 9459-9474.

Li, H. et al. (2023). Teach LLMs to personalise an approach for customizing LLMs. arXiv:2312.15503.

Liu, N. F. et al. (2024). Lost in the middle: How language models use long contexts. TACL, 12, 157-173.

Novak, I., Horvath, H., and Garcia, E. (2024). Cross-modal representation learning for generative intelligence. Journal of Generative Intelligence, 2024(1), 41-48.

Novak, E., and Silva, L. (2025). Consent-aware data processing frameworks for ethical AI systems. Journal of Responsible AI and Ethics, 2025(3), 42-49.

Nowak, L., and Petrov, A. (2025). Unified architectures for multimodal content generation. Journal of Generative Intelligence, 2025(1), 1-8.

Packer, C. et al. (2023). MemGPT: Towards LLMs as operating systems. arXiv:2310.08560.

Popescu, L., Horvath, M., and Novak, I. (2024). Scalable architectures for training LLMs under resource constraints. Journal of Generative Intelligence, 2024(1), 1-8.

Touvron, H. et al. (2023). LLaMA 2: Open foundation and fine-tuned chat models. arXiv:2307.09288.

Zhang, Y. et al. (2022). Personalised language models for stylistic generation. ACL 2022 Findings.

Zhong, W. et al. (2023). MemoryBank: Enhancing large language models with long-term memory. AAAI 2024.

Brown, T. et al. (2020). Language models are few-shot learners. NeurIPS, 33, 1877-1901.

Vaswani, A. et al. (2017). Attention is all you need. NeurIPS, 30.

Lindberg, S., Hansen, A., and Bianchi, A. (2024). Multimodal generative models for text-image-audio integration. Journal of Generative Intelligence, 2024(1), 33-40.

Costa, O., Schmidt, I., and Costa, L. (2024). Continual learning frameworks for long-lived generative AI systems. Journal of Generative Intelligence, 2024(1), 17-24.

Doshi-Velez, F., and Kim, B. (2017). Towards a rigorous science of interpretable machine learning.

arXiv:1702.08608.

Radford, A. et al. (2021). Learning transferable visual models from natural language supervision. ICML 2021.

Declarations

Funding

This research was supported by the French National Research Agency (ANR) under grant ANR-24-CE23-0046 (CAMG: Context-Aware Multimodal Generation) and the Austrian Science Fund (FWF) under grant P 42218-N. The funders had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

The CAMG implementation code, CAGQ evaluation framework, and anonymised user study response data are available from the corresponding author upon reasonable request. User study data are anonymised per GDPR requirements.

Ethical Approval

The user study received ethical approval from the Advanced Computing University Ethics Committee (reference ACU-ML-2024-028) and the Central European Tech University Ethics Committee (reference CETU-SD-2024-061). All participants provided informed written consent. Participation was voluntary with right to withdraw at any time.

ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437 Vol: 2025, Issues: 1, 2025

Appendix A

CAMG Implementation Details and CAGQ Evaluation Framework

The following provides CAMG component implementation details and the CAGQ evaluation dimension specifications.

Hierarchical Context Memory (HCM) Implementation

Multimodal Context Encoder (MCE) Configuration

CAGQ Evaluation Framework -- Five Dimensions