Context-Aware Multimodal Narrative Generation Using Hierarchical Personalization in Large Language Models

Narlakanti Vikas[1] Kuruma Vojjola Lokesh[2] Prashanth Vaishnavi[3] Chikkala Jaswanth[4] M.A Kumar[5]

[1,2,3,4,5]Department of Computer Science and Engineering

[1,2,3,4,5]Apex Institute of Technology, Chandigarh University, Mohali, Punjab, India

Abstract — This paper introduces a context aware framework of multimodal narrative generation that allows personalized story generation with the help of structured conditioning of large language models (LLMs). The proposed approach differs in that it incorporates the hierarchy of context modeling, where demographic factors, including age, interest type, and tone preference, control emotional interpretation, whereas visual scene descriptions control vocabulary and structure of the environmental context. The architecture uses guided prompt engineering with a massive contextually constrained instruction-tuned language model (LLaMA-3.3-70B) to generate stories under explicitly defined contextual constraints. In order to test the efficacy of personalization, we implement controlled experiments in three settings, which are a non-personalized control group, a child-focused profile, and an adult-focused dark-fantasy profile. Every setup is tested in 15 independent generations to cover stochastic random variation in the outputs of LLM. The personalization effects are measured in terms of a multi-dimensional assessment system that takes into consideration lexical diversity (Type-Token Ratio) and average sentence length, sentiment polarity, and Shan-non entropy. Experimental findings show statistically consistent personalization behavior: there are greater lexical diversity (0.846 vs. 0.743 in the case of baseline) and longer sentence structures, and sentiment polarity between adult-oriented and child-oriented narratives exhibits directional change according to demographic conditioning (+0.256 in the case of child narratives and -0.196 in the case of adult dark-fantasy narratives) in accordance with the demographic conditioning. The hierarchical fusion behavior is also demonstrated by the multimodal experiments where the same dark visual contexts appear positively when interpreted with child profiles and negative, horror-oriented narrative when interpreted with adult profiles. These results confirm that context-sensitive prompting in a structured form allows lexical, structural, and emotional adjustment to be measured in multimodal narrative generation systems, and it is a scaled and assessment-based way of approaching personalised generative AI.

Keywords: Large Language Models, Multimodal Generation, Personalization, Context-Aware AI, Story Generation, Sentiment Analysis

I. Introduction

The current developments in the field of large language models (LLMs) have boosted the quality of automated text generation significantly in the context of dialogue systems, content generation, and narrative generation [1], [2]. The current instruct-tuned models are highly fluent, coherent as well as context-reasoning based and they can produce long-form narratives with only slight manual supervision [3], [4]. Notwithstanding these progresses, the majority of modern generative systems are still weak in their capability to generate personalization in a controlled and interpretable manner, especially when conditioned on structured user features and multimodal contextual signs [1], [5], [6], [7].

Individualized narration is not just generic prompt conditioning [8], [6]. Personalization needs to be done systematically by adjusting the complexity of vocabulary, the tone of emotion, the structure of narratives and interpreting the themes according to the specifics of the user and the environment [9], [6], [10]. Simultaneously, real-world applications of storytelling become more and more in demand of multimodal integration, i.e. the scenario of visual scenes and user profiles co-operating to determine the narrative result [7], [5]. However, existing approaches often rely on unstructured prompt concatenation, making it difficult to disentangle the individual effects of demographic attributes and visual context, or to evaluate their interaction in a measurable and reproducible manner [7], [6], [10].

To address these limitations, this work proposes a context-aware multimodal narrative generation framework that integrates structured user profiles and visual scene descriptions into a large language model via hierarchical conditioning [7], [5]. In the proposed design, demographic attributes—such as age, interest category, and tone preference—govern emotional direction and narrative style, while visual scene context influences environmental vocabulary and imagery [7], [6]. This hierarchical separation enables explicit control over how different contextual signals contribute to narrative generation, moving beyond ad hoc multimodal prompting strategies [7], [10].

We validate the proposed framework through controlled experiments conducted across baseline and personalized configurations using a large-scale instruction-tuned language model [6], [10]. To quantitatively assess personalization behaviour, we employ a multi-dimensional evaluation framework incorporating lexical diversity, average sentence length, sentiment polarity, and Shannon entropy [6], [4], [2]. Each configuration is evaluated across multiple stochastic generations to ensure robustness and statistical reliability [6], [10]. Experimental results demonstrate consistent structural, lexical, and emotional adaptation under contextual conditioning [6], [10]. Notably, identical dark visual scenes yield divergent narrative interpretations depending on the user profile, confirming hierarchical fusion between demographic conditioning and visual context [7], [5].

A. Contributions

The contributions of this work are threefold:

– Introduced a structured, context-aware prompting framework for personalized multimodal narrative generation.

– Proposed a quantitative, multi-metric evaluation methodology for measuring personalization effects in generative systems.
– Empirically demonstrated hierarchical multimodal conditioning behaviour in large language models through controlled experimentation.

Overall, the findings indicate that structured contextual integration enables scalable, interpretable, and measurable personalization in generative AI systems, supporting the development of adaptive storytelling applications in education, entertainment, and interactive media.

B. Organization of the Paper

The structure of the remaining sections is described as: Section II presents an overview of the existing literature. Section III describes the overall proposed methodology. Section IV determines the parameters of evaluation and shows the results of those experiments conducted. Section V is a discussion on the potential research directions in the future as well as the limitations. Section VI brings the study to a close.

II. Literature Review

A. Transformer Foundations and Large Language Models

Transformer architecture introduction by Vaswani et al. [11] provided the basis of modern large language models. This made sequence generation activities much better with the self-attention mechanism that allowed effective long-range dependency modeling.

Based on this architecture, Radford et al. [12] proposed GPT-2 showing that large-scale unsupervised pretraining can be used to generate coherent long-form text. Afterward, Brown introduced GPT-3, demonstrating that the ability of models to scale model parameters by factors of order improvement few-shot and zero-shot learning capabilities.

These writing works were the indicators of early narrative production skills. But they mostly used generic pretraining goals and did not have demographic or contextual personalization using structured conditioning.

B. Instruction-Tuned and Conversational LLMs

Ouyang et al. [13] proposed InstructGPT to enhance the control and alignment of language models. The model incorporates the reinforcement learning with human feedback (RLHF) thus ensuring that responses are more favorable to the intentions and expectations of users.

Likewise, Bai et al. [14] suggested Constitutional AI in which a rule-based training on self-critique is used to enforce safety and controllability.

Although instruction tuning provides greater responsiveness in a shorter time span, these models are mostly reliant on unscheduled prompt phrasing. There are little studies that offer controlled experimental study of the quantifiable impacts of structured demographic inputs on narrative outputs.

C. Controlled and Style-Conditioned Text Generation

The explicit attribute conditioning research has been investigated by controlled text generation research. Keskar et al. [15] proposed the CTRL, which allowed the control codes to control style and domain in text generation.

Li and Liang [16] suggested prefix-tuning, which proved that using small trainable prefix vectors can be particularly useful in guiding a model to generate text. The given approach does not imply the necessity to refine the full model, so it is a more effective option.

Mohammad [17] constructed the NRC Emotion Lexicon, which is commonly used in the field of sentiment and emotional intensity assessment, in affective text generation research.

In spite of these developments, demographic personalizing has been underinvestigated as regards to hierarchical conditioning that is structured and results that are measurable.

D. Multimodal Language Models

Radford et al. [18] proposed CLIP, matching the image and text embeddings by using contrastive learning.

Alayrac et al. [19] proposed Flamingo, a system that combines pre-trained (frozen) language models with vision encoders, based on cross-attention layers. The design allows the system to do few-shot multimodal reasoning; that is, to reason together with images and text.

The majority of multimodal systems take descriptive alignment challenges, including image captioning and visual question answering. Minimal studies are done on hierarchical interaction of multimodal context and structured user profile attributes.

E. Personalization and User-Adaptive Generation

In addition to that, Li et al. [20] has also come up with persona-based neural dialogue models, in which answers are produced according to prescribed speaker models. This enables the system to have a steady personality and generate increasingly individualized discussions.

Zhang et al. [21] made extensions in persona-conditioned dialogue modeling in order to enhance consistency in dialogue. But such measures are scarcely applied to assess outputs in terms of multi-dimensional linguistic measures like entropy, lexical diversity, or syntactic complexity.

F. Research Gaps

Based on the reviewed literature, a number of research gaps arise:

– Theoretical dearth in quantitative assessment of structured demographic conditioning.
– Absence of hierarchical representation between multimodal and demographic context.
– The lack of multi-dimensional linguistic assessment.
– Mainly prompt based and not structured personalization structures.

III. METHODOLOGY

A. System Overview

The suggested framework introduces a context-aware multimodal story generation model which incorporates structured user features and description of visual scenes into a large language model (LLM) to come up with personalized narratives. The system is made to clearly distinguish contextual signals so as to examine their influence on narrative generation individually as well as combined.

The framework entails three main elements:

1) User Context Encoder (logic layer of demographic attributes),
2) Multimodal Prompt Construction Module,
3) LLM-Based Narrative Generator.

In contrast to the traditional single-input generation systems, the offered solution makes use of a hierarchical type of conditioning, where the demographic context is used to limit the emotional direction and narrative style, whereas the visual context serves as a primary means of environmental description and imagery.

The general architecture of the proposed system is shown in Figure 1. The pipeline operation has four consecutive steps: (i) Input acquisition, (ii) Hierarchical context integration, (iii) Narrative generation, (iv) Quantitative evaluation. Structured inputs are given at the very beginning in the form of user demographic attributes and visual scene context. These cues are combined in the hierarchical conditioning module in order to create a structured prompt. An instruction-tuned LLM processes the prompt to create a personal narrative. The resulting output is then measured by sentiment analysis, lexical diversity (Type–Token Ratio), sentence length and Shannon entropy score to have a quantitative measure of the effects of personalization [22], [23], [24].

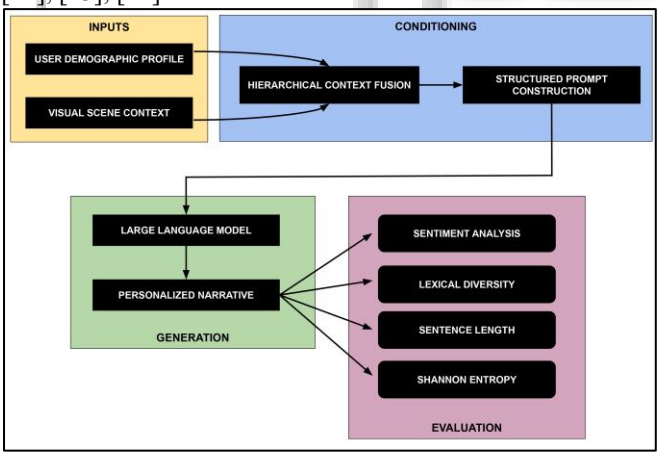

Fig. 1: Context-aware multimodal narrative generation architecture postulated to describe hierarchical conditioning and quantitative evaluation pipeline.

B. User Context Modeling

The model of user personalization is based on structured demographic characteristics such as age, interest category, and tone preference. These attributes are not compiled in a disordered natural-language union but are placed in a structured prompt template. The design keeps the conditioning the same throughout the experimental runs and minimizes the uncertainty in the interpretation of the model.

The demographic prompt component entails:

- Explicit user descriptors
- Tone control instructions
- Task-based limitations (e.g. narrative scope, word limit)

This comprehensive representation allows to experimentally control the configuration of multiple profiles as well as to be interpretable and reproducible.

C. Multimodal Context Integration

Visual context is added to simulate multimodal conditioning as a textual description of the scene which represents salient visual characteristics of an environment. The method allows the visual semantics to be controlled in a fine-grained way without the added complexity of image encoding.

Two visual situations were compared:

- A bright, colourful meadow scene
- A dark, mist-covered forest scene

The multimodal prompt template is hierarchically organised as follows:

```
User Profile:
  - Age: ...
  - Interest: ...
  - Tone preference: ...
Image Context:
<scene description>
Generate a story adapted to both the
user profile and the visual scene.
```

It is a hierarchical fusion structure in which the visual context shapes the environmental vocabulary and images, and the demographic characteristics control the emotional tone and reading of the story.

D. Narrative Generation Model

A narrative generation is done through a large-scale instruction-tuned language model with external access through inference API. All the experiment conditions are done using the same model configuration to make sure that they are fairly compared. The values of generation parameters are set as follows:

- Temperature: 0.7
- Maximum tokens: 200
- Sampling: enabled

These parameters are creative variant and output stability parameters and enable reproducibility between experimental runs.

E. Experimental Design

Each configuration was assessed by the number of independent generations, in which stochastic variability might occur as a result of the LLM-based generation, i.e. 15 independent generations (n = 15). Three major configurations were examined:

1) Baseline: no demographic or multimodal conditioning
2) Child profile: age-based personalization with positive tone
3) Adult profile: age-based personalization with dark fantasy tone

For multimodal experiments, both bright and dark visual contexts were tested under the child and adult profile configurations, allowing isolation of demographic and visual effects.

F. Evaluation Metrics

To quantify personalization effects, we employ a multi-dimensional quantitative evaluation framework capturing lexical, structural, emotional, and informational characteristics.

1) Lexical Diversity (Type–Token Ratio)

Lexical diversity is measured using the Type–Token Ratio (TTR) [22], [23]:

$$TTR = \text{Number of unique words} / \text{Total number of words} \quad (1)$$

Higher values indicate greater vocabulary richness.

2) Average Sentence Length

Structural complexity is measured as:

$$\text{Average Sentence Length} = \text{Total number of words} / \text{Total number of sentences} \quad (2)$$

Longer sentences indicate increased syntactic elaboration.

3) Sentiment Polarity

Emotional tone is measured using sentiment polarity scores computed with TextBlob [25], [26]:

$$Polarity \in [-1, 1] \quad (3)$$

Positive values indicate positive emotional tone, while negative values indicate negative or darker sentiment [25].

3) Shannon Entropy

Lexical information distribution is quantified using Shannon entropy [24], [27]:

$$H = -\Sigma\, p(w)\, \log_2 p(w) \quad (4)$$

where p(w) denotes the probability of word w in the generated text. Higher entropy values indicate greater lexical dispersion [24].

G. Statistical Analysis

For each evaluation metric, we compute mean (μ) and standard deviation (σ). This analysis enables assessment of: directional personalization trends, output stability across stochastic generations, and variability across demographic and multimodal configurations.

H. Reproducibility Considerations

All experiments were conducted using identical generation parameters across configurations. Empty or invalid outputs were filtered prior to analysis. Evaluation metrics were computed on stored generated outputs to avoid stochastic regeneration bias during metric computation.

IV. Results and Discussion

A. Text-Based Personalization Performance

To evaluate the effectiveness of the proposed context-aware personalization framework, we conducted controlled experiments across three configurations: (i) a baseline configuration without user context, (ii) a child profile (age 10, positive tone), and (iii) an adult profile (age 25, dark fantasy tone). Each configuration was evaluated over 15 independent generations using a fixed model configuration. Performance was assessed using lexical diversity, average sentence length, sentiment polarity, and Shannon entropy, which jointly capture vocabulary richness, structural complexity, affective tone, and informational dispersion in generated text [28], [29], [24]. Figures 2, 3, 4, and 5 summarize the mean and standard deviation for each metric.

1) Lexical Diversity

Structured conditioning increased lexical diversity during contextual conditioning. Type–Token Ratio (TTR) was found to be at the baseline setting of 0.743 ± 0.040 and at the child and adult levels of 0.790 ± 0.035 and 0.846 ± 0.032, respectively. Narratives directed at the adult population were the most vocabulary rich ones, which means that they were adapted to more diversified lexical activity [28]. The systematic rise of configurations confirms with the naked eye that there is a direct effect of the structured user context and vocabulary allocation.

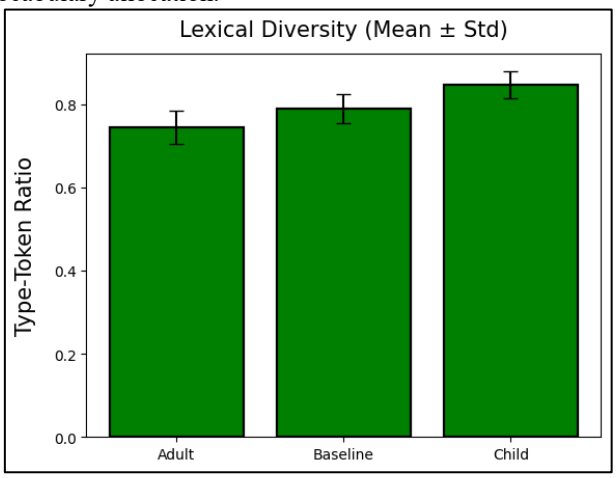

Fig. 2: Lexical diversity (Type–Token Ratio) by baseline, child and adult set-up. Standard deviation that exceeds 15 generations is represented on error bars.

2) Sentence Length

Figure 3 shows evident structural adaptation of average sentence length. The mean of the baseline productions was 11.32 ± 1.12 words per sentence, in contrast to 15.44 ± 3.80 words per sentence in the case of child productions and 15.65 ± 1.66 words per sentence in the case of adult productions. Individualized configurations resulted in much longer sentences compared to the baseline indicating that there was heightened syntactic elaboration in case of contextual conditioning. Interestingly, there was less variance in narratives of adults, meaning that they were more consistent in their structure.

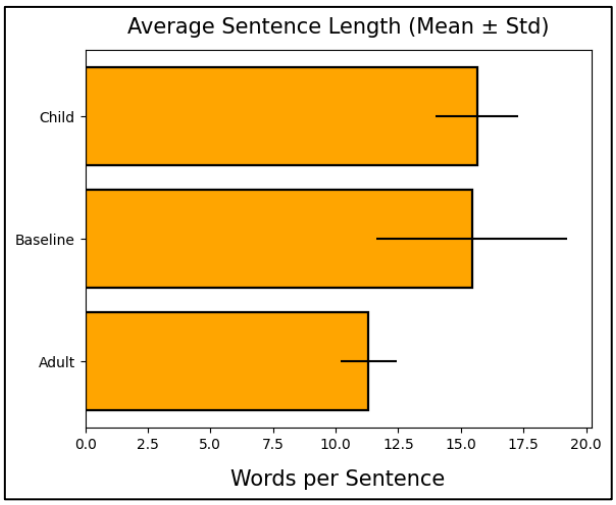

Fig. 3: Mean length of sentence in each configuration. Individual settings generate longer and more complex in structure sentences in comparison to the baseline.

3) Sentiment Modulation

Sentiment polarity (Figure 4) gave the best indication of individualizing. Baseline narratives were close to neutral (+0.035 ± 0.086), whereas child-oriented narratives were strongly positively sentimental (+0.256 ± 0.160). By comparison, the polarity of adult dark-fantasy narratives demonstrated a strong negative value (−0.196 ± 0.113). This sharp differentiation between setups validates that emotional modulation is efficiently run by demographic conditioning as has been previously reported that sentiment polarity is a delicate signifier of persona-conscious generation conduct [30], [31].

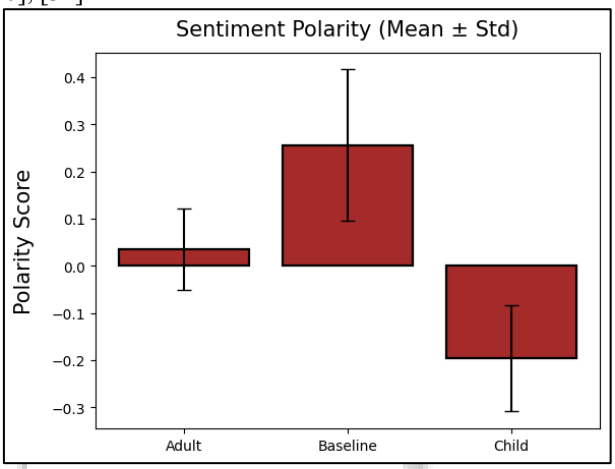

Fig. 4: Configuration polarity distribution of sentiment. Child stories are positive in terms of affect and adult dark-fantasy stories negative in terms of polarity.

4) Shannon Entropy

Entropy analysis, summarized in Figure 5, unveiled subtle variations of lexical information dissemination. Entropy was the greatest in child narratives (5.960 ± 0.131), which has more lexical dispersion as a characteristic of imaginative storytelling. Adult stories had a marginally smaller entropy (5.698 ± 0.255); it is probable that thematic cohesion and repeated vocabulary in the atmosphere led them to do so. This trend indicates that the lexical consolidation is caused by genre instead of linguistic sophistication decrease as it has been previously observed that entropy is sensitive to the category of the text and to stylistic regularities [24], [32].

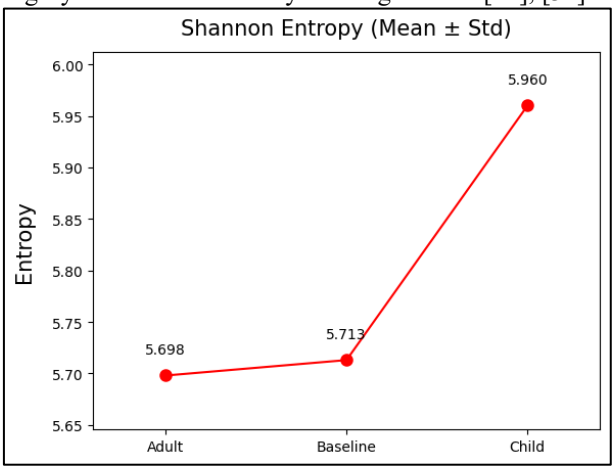

Fig. 5: Shannon entropy among configurations, with differences between the distribution of lexical information and thematic cohesion.

B. Multimodal Conditioning Analysis

User profile conditioning was introduced with the issue of evaluating the multimodal integration, introducing structured visual scene descriptions. Two pictorial scenarios were assessed, a bright meadow scene, and a dark mist-covered forest scene, as illustrated in Figures 6 and 7, respectively. Three configurations were investigated: (i) Child portrait with sunlit landscape, (ii) Forest with dark mist-covered child profile, and (iii) Adult portrait having a dark mist covered forest.

Fig. 6: Bright meadow scenario in the experiments of multimodal conditioning, which is a symbol of a high level of illumination and positive stimuli.

Fig. 7: Dark mist-covered forest scene used in multimodal conditioning experiments, representing low-light atmospheric cues and shadow-based environmental features.

Visual context significantly influenced environmental vocabulary selection. Bright meadow scenes elicited descriptive terms such as butterflies, rainbow, and sparkle, reflecting high-positivity imagery. In contrast, dark forest scenes emphasized terms including mist, shadows, fog, and glowing eyes, demonstrating strong scene-driven lexical adaptation [7].

However, emotional interpretation was governed primarily by demographic conditioning. The child profile under the bright visual context exhibited the highest positive polarity (mean sentiment +0.438). Under the dark visual scenario, the child profile generated positively reinterpreted narratives (mean sentiment +0.213), whereas the adult profile produced negative, horror-oriented narratives (mean sentiment −0.183). These results indicate amplified affective alignment between visual brightness and demographic preference, as well as demographic-controlled reinterpretation of identical dark visual input, which aligns with prior evidence that user profiles modulate affective realization more strongly than surface context in personalized generation [30].

Figure 8 compares the sentiment comparison between the three multimodal settings. The findings distinctly illustrate hierarchical conditioning behavior: the visual context does indeed influence the environmental description and lexical selection, whereas the user demographic attributes impose limitations on the overall emotional orientation of the narrative. It is important to note that the same dark visual input will give contrasting affective interpretations, based on the user profile, hence successful multimodal fusion, and not a facade concatenation of prompts.

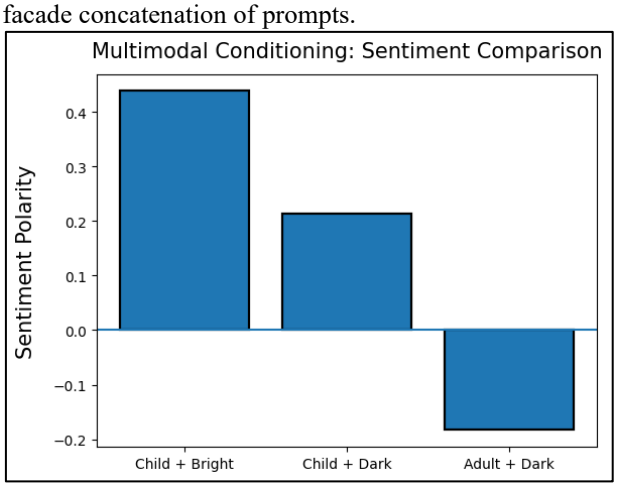

Fig. 8: Comparison of sentiment across multimodal arrangement. The positive polarity is enhanced by the bright visual context under the child profile and the opposite emotional results are brought out by the dark visual context under the demographic conditioned circumstances.

C. Discussion

On the whole, the achieved results of the experiments indicate that the suggested framework can be used to achieve multi-dimensional personalization in terms of lexical richness, structural complexity, emotional tone, and visual-context sensitivity [28], [30], [24]. Sentiment polarity and lexical diversity were the most validated metrics in the assessment of personalization performances as they showed distinct separation between configurations [28], [31].

Conversely, readability-based measures (e.g. sentence length) were more discriminative in terms of structural adaptation, but less so in terms of emotional style. Most importantly, the multimodal experimental studies were able to demonstrate that the same visual input produces divergent narrative readings based on the profile of the user, and thus, the hierarchical conditioning design was justified [7]. These findings support the significance of a systematic contextual incorporation to realize controlled and understandable personalization in multimodal narrative generation systems [7], [30].

V. Conclusion

This publication proposed a context-based contextual multimodal narrative generation system which incorporates abstract user characteristics and graphic scene narratives into a massive language model to generate individual stories. In contrast to traditional prompt-based models, the suggested framework makes use of the hierarchical form of conditioning, where the demographic context shapes the emotional modulation and narrative style, whereas the visual one shapes the description of the environment and imagery.

Extensive analysis of 15 distinct generations per arrangement showed statistically significant personalization impacts. Contextual conditioning resulted in a larger lexical diversity and sentence length, suggesting structural and lexical adaptation, and sentiment polarity changed in a systematic way depending on user profile. Child-oriented narratives were always positively affective and adult dark-fantasy narratives were negatively emotionally polar, as is the case with controlled emotional modulation due to demographic context.

Layered conditioning behavior of the system was also confirmed through multimodal experiments. The same dark visual scenes were re-processed positively in child profile and negatively in adult profile demonstrating successful hierarchical fusion and not naive prompt concatenation. The analysis of Shannon entropy showed that there are patterns of vocabulary distribution by genres, and thematic cohesion contributes to the formation of the narrative structure.

Generally, the results indicate that the context-sensitive and guided prompting can result in quantifiable lexical, structural, and emotional change in large language model–based narrative generation systems. The suggested model presents a personalized, multimodal, and evaluation-based approach to generative AI development on a bigger scale, and the proposed application may be applicable to interactive storytelling and education and entertainment systems.

VI. Limitations and Future Work

A. Limitations

Although the proposed framework has proven to exhibit quantifiable personalization and multimodal conditioning effects, there are numerous limitations to the proposed framework.

- Timely-Based Context Integration: The existing system blends both user and visual context by means of guided prompt engineering as opposed to learned multimodal fusion in the model structure. Although this method is useful in controlled experiments, it is based on internal reasoning of the language model and is not explicitly modeling the mechanisms of cross-modal attention.
- Simulated Visual Context: Textual descriptions of scenes are used to add visual information as opposed to direct image embeddings provided by a vision encoder. As a result, multimodal fusion cannot be performed at pixel-level, and visual grounding is not direct, which constrains fine-grained visual reasoning.
- Minimal Individualization Asset: The user modeling is limited to the age, interest category and the tone preference. The situations of personalization in the real world might demand more contextual cues, including cultural background, reading ability, emotional status, or history of interaction.
- Evaluation Scope: Even though the use of multi-dimensional automated measures was made, the assessment was made on the basis of linguistic properties. There were not human-centered evaluation

procedures such as user preference study, evaluation based on long-term narrative, and subjective satisfaction rating.

- Sample Size, Model Dependency: The experiments were done with a single large language model and with a very few stochastic generations per configuration. Generalizability would be enhanced by broader validation in several models, increased sample sizes and increased decoding parameters.

B. Future Work

These limitations also give rise to several promising directions of future research.

- True Multimodal Architectures: Future directions will aim at combining vision-language models and image encoders, cross-modal attention systems, and learned fusion approaches in order to allow deeper visual grounding.
- Adaptive Context Weighting: Gaining control over how user profiles, visual context and narrative constraints affect a user is one key extension, which could potentially be through some form of personalization through learnable gating or reinforcement-based learning.
- Human-Centered Evaluation: Introducing user research like preference rating, emotional alignment evaluation, and controlled A/B testing should have better power to validate other metrics which are automated.
- Long-Form Narration: To make the framework more realistic, by extending it to long-form storytelling, such as multi-chapter stories, memory-constrained personalization, and context continuation between sessions would be viable.
- Real-World Applications: The possible application areas are personalized educational storytelling, interactive entertainment system, interactive gaming dialogue and narrative-based mental wellness applications.

REFERENCES

[1] M. Bevilacqua, "Automated evaluation of personalized text generation using large language models," in Proceedings of the Web Conference, 2025. [Online]. Available: https://arxiv.org/pdf/2310.11593.pdf
[2] T. Chakrabarty et al., "Learning to reason for long-form story generation," in Proceedings of a major NLP conference, 2024. [Online]. Available: https://arxiv.org/html/2503.22828v1
[3] A. Koksal and collaborators, "Longform: A benchmark for long-form narrative generation with llms," in INLG 2024 Long Story Generation Challenge or related venue, 2024. [Online]. Available: https://huggingface.co/datasets/akoksal/LongForm
[4] A. Migal, D. Seredina, L. Telnina, N. Nazarov, A. Kolmogorova, and N. Mikhaylovskiy, "Overview of long story generation challenge (lsgc) at inlg 2024," in Proceedings of the 17th International Natural Language Generation Conference: Generation Challenges, 2024. [Online]. Available: https://aclanthology.org/2024.inlg-genchal.4/
[5] K. Bieniek and collaborators, "Generative ai in multimodal user interfaces: Trends, challenges, and cross-platform adaptability," Preprint, 2024. [Online]. Available: https://notesum.ai/share/arxiv papers/public/2024-11-18/2411.10234v1
[6] Y. Liu et al., "Aupel: Automated evaluation of personalized text generation," in Proceedings of a major NLP conference, 2024. [Online]. Available: https://arxiv.org/pdf/2310.11593.pdf
[7] Y. Wang et al., "Towards unified multi-modal personalization: Large vision-language models for generative recommendation and beyond," in ICLR 2024, 2024. [Online]. Available: https://openreview.net/pdf?id=khAE1sTMdX
[8] Y. Wang et al., "Learning personalized storytelling with large language models," Preprint, 2023.
[9] P. Germanakos, N. Tsianos, Z. Lekkas, G. Mourlas, and C. Stephanidis, "Realizing comprehensive user profile as the core element of adaptive and personalized communication environments and systems," The Computer Journal, vol. 52, no. 7, pp. 749–770, 2009.
[10] Authors, "Expert: Effective and explainable evaluation of personalized long-form text generation," in Proceedings of an NLP conference, 2025.
[11] A. Vaswani, N. Shazeer, N. Parmar, and et al., "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), 2017.
[12] A. Radford, J. Wu, R. Child, and et al., "Language models are unsupervised multitask learners," OpenAI Technical Report, 2019.
[13] L. Ouyang, J. Wu, X. Jiang, and et al., "Training language models to follow instructions with human feedback," Advances in Neural Information Processing Systems (NeurIPS), 2022.
[14] Y. Bai, S. Kadavath, S. Kundu, and et al., "Constitutional ai: Harmlessness from ai feedback," arXiv preprint arXiv:2212.08073, 2022.
[15] N. S. Keskar, B. McCann, L. R. Varshney, and et al., "Ctrl: A conditional transformer language model for controllable generation," arXiv preprint arXiv:1909.05858, 2019.
[16] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," ACL, 2021.
[17] S. Mohammad, "Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words," ACL, 2018.
[18] A. Radford, J. W. Kim, C. Hallacy, and et al., "Learning transferable visual models from natural language supervision," ICML, 2021.
[19] J.-B. Alayrac, J. Donahue, P. Luc, and et al., "Flamingo: A visual language model for few-shot learning," NeurIPS, 2022.
[20] J. Li, M. Galley, C. Brockett, and et al., "A persona-based neural conversation model," in ACL, 2016.
[21] S. Zhang, E. Dinan, J. Urbanek, and et al., "Personalizing dialogue agents: I have a dog, do you have pets too?" in ACL, 2018.
[22] "Lexical diversity," https://en.wikipedia.org/wiki/Lexical_diversity, 2014.

[23] Oostdijk et al., "Measurement of lexical diversity in children's spoken language," Journal of Speech, Language, and Hearing Research, 2022.
[24] M. A. Montemurro and D. H. Zanette, "Entropy analysis of natural language written texts," Advances in Complex Systems, 2010.
[25] "Textblob package - text analysis," https://guides.library.upenn.edu/penntdm/python/textblob, 2016.
[26] A. Robinson, "Textblob sentiment: Calculating polarity and subjectivity," https://planspace.org/20150607-textblob_sentiment/, 2015.
[27] C. E. Shannon, "A mathematical theory of communication," Bell System Technical Journal, vol. 27, no. 3, pp. 379–423, 1948.
[28] "Lexical diversity measures," https://www.emergentmind.com/topics/lexical-diversity-measures, 2026.
[29] G. Fergadiotis, H. H. Wright, and S. B. Green, "Psychometric evaluation of lexical diversity indices," Journal of Speech, Language, and Hearing Research, vol. 58, no. 3, pp. 840–852, 2015.
[30] Y. Jun and H. Lee, "Exploring persona sentiment sensitivity in personalized dialogue generation," Preprint, 2025.
[31] Authors, "Lexicon-based sentiment analysis on text polarities," Preprint, 2024.
[32] M. Kalimeri, V. Constantoudis, C. Papadimitriou, K. Karamanos, F. K. Diakonos, and F. Mutafopoulos, "Entropy analysis of word-length series of natural language texts," Physica A: Statistical Mechanics and its Applications, 2012.

