

UGC-Driven Dynamic Evolution of Public Art Installations via Multimodal Transformer: A Tourist Content Perspective

Junjun Jiang

School of Art and Design, Jiangxi Institute of Technology, Nanchang 330098, Jiangxi, China

Corresponding author: Junjun Jiang (e-mail: 13576279627@163.com)

ABSTRACT Public art installations have long been limited by their inability to perceive and respond in real time to visitor-generated content in situ. To address this limitation, this study proposes a dynamically evolving public art installation system driven by tourist user-generated content and supported by a multimodal Transformer architecture. The system processes heterogeneous image, text, and audio inputs through a hierarchical cross-modal attention mechanism to extract collective affective semantics from visitor-generated signals. A differentiable parameter-mapping network then converts these semantic representations into real-time control actions for installation color, morphology, and rhythmic behavior. The framework also incorporates edge-side deployment, lowlatency communication, and continuous data-stream processing, which are essential for public environments involving dense wireless signal propagation and multimodal sensing. Three field tests were conducted in urban public plazas in China. The results show that tri-modal semantic understanding achieved a fusion accuracy of 91.2%, while end-to-end latency remained below 500 ms. Compared with static installations, the proposed system increased average visitor dwell time by approximately 100.2% and secondary user-generated content publication by 121.8%. During 30 days of continuous operation, system stability reached 99.2%. These results demonstrate the feasibility of integrating multimodal semantic understanding, real-time signal acquisition, and responsive visual control for human-machine co-creative public artworks, and provide an engineering pathway for interactive installations operating in complex wireless and acoustic environments.

INDEX TERMS multimodal transformer, public art installation, user-generated content, cross-modal attention, signal propagation

I. INTRODUCTION

The convergence of digital technology and public art shows that a pressing research question has emerged. Moreover, the significant challenge of converting massive volumes of tourist usergenerated content (UGC) into semantic signals that drive real-time installation evolution indicates that this area remains largely unresolved. Furthermore, the findings appear to demonstrate that this question sits at a critical intersection of smart city development and interactive art.

In light of these results, traditional public art installations shows that pre-programmed routines produce static or cyclically repeating visual effects with no capacity to perceive visitors' dynamic emotional states or culturally diverse expressions. Research shows human-installation interaction shallow – confined to simple trigger-response mechanisms, far from genuine content resonance. However, the significant advances in deep learning indicates that new possibilities have emerged for addressing these limitations. Additionally, Vision Transformer [1] shows that substantial gains in

visual representation are achievable through its global self-attention modeling mechanism.

Given that evidence demonstrates explorations into quantum-classical hybrid neural networks have further extended the theoretical boundaries of visual transfer learning [2], the results indicates that these developments establish a key methodological foundation for multimodal content perception in complex public environments. Findings show vision-language pre-training frameworks achieve precise image-text semantic alignment. Nevertheless, the significant evidence shows that coordinated breadth-and-depth attention mechanisms [3] could demonstrate meaningful improvements in capturing fine-grained affective content. Furthermore, the findings appear to indicate that report refinement strategies targeting key semantic knowledge [4] support these improvements.

In light of the evidence, graph-structured fashion semantic knowledge modeling [5] could demonstrate that structured prior knowledge enhances the quality of vision-language

representations. Evidence shows multimodal audio-visual perception robotic systems integrating Transformer and generative adversarial networks [6] support real-time artistic response. However, the results shows that UGC-driven research for public art installations indicates that this area remains at an early stage. Moreover, the significant challenges that persist in open environments indicates that filtering anomalous semantics from noisy UGC is technically demanding. Thus, the evidence appears to demonstrate that sustaining stable long-term system operation introduces additional engineering constraints.

Given that findings suggest maintaining diversity and coherence in evolutionary outputs under real-time latency requirements remains an open problem [7], these factors indicates that practical deployment continues to face important limitations. Research shows these gaps necessitate new approaches. To address these challenges, we propose an adaptive framework that leverages tourist-generated content to stimulate the organic growth of public art. By utilizing a transformer-based multimodal architecture, the system optimizes artistic interaction through automated semantic extraction and responsive visual adjustment, replacing the static paradigms of the past.

Furthermore, the significant system indicates that heterogeneous multi-source UGC – image, text, and audio – provides the key input for a cross-modal attention fusion module. In light of the evidence, the results appear to suggest that this module extracts high-dimensional semantic representations of collective visitor affect. Findings show a differentiable parameter mapping network translates these representations into physical visual parameters. However, the significant findings shows that color, morphology, and rhythm parameters could demonstrate that this approach closes a complete perception-understanding-evolution loop. Therefore, the study indicates that equipping public art installations with autonomous evolutionary capacity appears to represent a critical advancement. Given that the evidence demonstrates that the system actively perceives visitor affect and continuously generates differentiated visual outputs, the results indicates that this provides both a systematic theoretical basis and a reproducible technical pathway. Moreover, the significant findings shows that this approach could demonstrate a new paradigm of humanmachine co-creative art in public space.

II. METHODOLOGY

A. MULTIMODAL TRANSFORMER ARCHITECTURE FOR UGC CONTENT UNDERSTANDING

In the designed architecture there will be three kinds of input – image, text and audio – taken out from tourist UGC data. And in public art environments especially there exist many difficulties: high level of content noise, unbalanced distribution of modalities etc. In order to solve above problems such an end-to-end cross-modal semantic understanding framework has been set up here [8]. As shown, firstly each kind of individual data will go through its own encoding process

respectively. Among them images will adopt the method of patch embedding to divide each input image into several sequences made of non-overlapping patches sized in 16×16 after which they need to carry on the positional encoding step before entering the Vision Transformer encoder, in total consisting of twelve layers which can not only help grasp certain kinds of fine-grained texture features but also understand some big-picture compositional structure.

In respect of the textual information, reviews made by visitors plus topic hashtags are processed via a BERT-based tokenizer. Afterwards bidirectional attention mechanisms help to infer their affective polarity and also their semantic intent. As far as audio is concerned, firstly the soundtrack of each short video is transformed into a corresponding mel-spectrogram. Then a lightweight ViT module - which shares its weights with the image branch - will encode the aforementioned acoustic features into corresponding audio embedding representations. Such a weight-sharing design can keep the total number of parameters under control but meanwhile is capable enough of projecting all the three kinds of modalities into one common representation space [9].

Concerning the cross-modal fusion, here it adopts a method based on hierarchical cross-attention mechanisms. In other words, text semantic vectors are taken as queries which then attend over both image patch feature sequences and audio frame feature sequences through multi-head cross-attention operations. The learnable modal weight gate controls each modality's share of contributions towards producing the fused representation. Hence it could reduce the effect brought by low-quality UGC examples over overall semantic comprehension; The architecture operates through a rigorous sequence: first, it encodes raw inputs into discrete embeddings; next, it aligns these features through a hierarchical crossattention gate; and finally, it projects the fused representation via two feed-forward layers into a 128-dimensional affective-semantic embedding space. This refined syntactic structure clarifies that the 'intrinsic correlation mechanism' is driven by the dynamic interplay between query-based attention and learnable modal weights. Such kind of semantic representation will work as drive signal to the subsequent installation parameter mapping network [10], besides inference latency also became one consideration factor in architecture design here. Model compression via knowledge distillation helped shrink the backbone size making it eligible for edge-side deployment; meanwhile no matter during model training phase or service running phase, end-to-end time cost – from UGC input to semantic output – was kept less than 500 milliseconds all the time, therefore ensuring system running smoothly and stably providing solid basis for real-time installation evolution.

B. UGC DATA COLLECTION, CLEANING, AND ANNOTATION

The data collection work was carried out at three representative urban public art plazas in China. It took both

online and offline ways to get data together. Online part adopted open APIs provided by Weibo, Xiaohongshu, and Douyin platforms. Taking one installation as a sample, a 500-meter radius of interest zone was built up according to its geographical coordinates. And then the relevant materials – texts, pictures, and short videos – which contained some location-related hashtags were obtained constantly over twelve months. There were a large quantity of raw data, totaling about 214,736.

Offline portion gained access with the help of the administration staff here. First of all, the four sets of directional microphone arrays and two wide-angle cameras were equipped in several directions close to the installations under research. Then they recorded ambient audio at 44.1 kHz and behavioral video at 25 fps [11]. Data cleaning was done following a three-stage process. Firstly, according to certain rules, samples of very low quality were eliminated, including pictures whose size is less than 224×224 pixels, texts which contain fewer than eight valid characters, and audios whose SNR (Signal-To-Noise Ratio) value is smaller than 15 dB. Those seriously blurred, completely empty, extremely noisy entries were also removed.

Then, via Perceptual Hash, nearly identical pictures were marked out and removed. Thirdly, employing a regular expression-based pipeline, commercial advertisement and irrelevant content were filtered out. Finally, there remained 67.3% of the whole, which formed the final main body dataset consisting of totally 144,457 multimodal records [12]. Annotations adopted three-level semantics labeling framework. On the first level, three trained raters gave every sample one of three affective valences - positive, negative, and neutral respectively through their own judgment independently. Majority voting method decided what kind of labels will eventually belong to this piece of sample, and Kappa coefficient measured inter-annotator agreement, reaching 0.76. Secondly, multi-label classification was performed on eight predefined public art interaction themes.

Crucially, while affective valences capture transient emotional states, themes like cultural identity are treated as high-level sociocognitive constructs within the content semantics, separate from the real-time affective valence (positive/negative/neutral) used for immediate installation rhythm adjustment. Thirdly, the interaction intensity was quantified by means of three social engagement metrics – likes, comments, and shares. They were finally added up and turned into one single normalized engagement score recorded as a continuous value.

C. DYNAMIC EVOLUTION DRIVING MECHANISM FOR PUBLIC ART INSTALLATIONS

The dynamic evolution mechanism takes multimodal semantic vectors as its core input. Moreover, the significant parameter mapping network shows that three functional modules work in sequence to form an end-to-end pipeline from abstract semantic representation to concrete installation state. Given that the evidence demonstrates that sequential

processing could indicate a structured design, the evolution scheduler and physical rendering layer appear to support this architectural approach.

Furthermore, the parameter mapping network shows that three fully connected layers project the 128-dimensional joint embedding vector into a 28-dimensional installation control parameter space through nonlinear mapping. The mechanism pipeline covers color temperature, deformation, frequency, and transparency. Drawing on the 'Affective Filter Hypothesis,' these conscious system modulations aim to lower the subconscious 'interaction barrier' for visitors. By dynamically thinning the affective filter through real-time rhythm adjustment, the installation facilitates a more seamless transition from passive observation to intuitive, subconscious engagement, ensuring that teacher-led design interventions effectively stimulate genuine emotional resonance, while the smoothing constraint shows that abrupt inter-frame discontinuities are prevented during mapping [14].

Additionally, the significant evolution scheduler shows that a sliding-window aggregation strategy computes an exponentially weighted average over semantic vectors from the preceding 180 seconds, thereby dampening the influence of short-term anomalous samples. In light of the findings on temporal aggregation, the results could demonstrate that an evolution trigger threshold appears to ensure that a new parameter update is initiated only when the cosine similarity between the current aggregated semantic vector and the previous evolution state falls below 0.82. Scheduler shows threshold balances responsiveness with visual stability [15]. Therefore, the critical physical rendering layer shows that control parameters are transmitted to an LED matrix controller and mechanical deformation drive units via the DMX512 protocol. Notwithstanding the technical complexity, the significant findings indicates that end-to-end rendering latency appears to have been measured at a stable 120 milliseconds, satisfying the realtime requirements for smooth visual evolution in public environments [16].

D. EXPERIMENTAL PLATFORM AND EVALUATION METRICS

The experimental platform demonstrates that its architecture spans three layers: hardware infrastructure, software systems, and evaluation data collection. Moreover, the significant findings shows that on-site deployment across the three urban public art plazas could indicate meaningful real-world applicability. Furthermore, the hardware infrastructure centers on two edge computing servers, each equipped with an NVIDIA A100 GPU, and the key evidence shows that these nodes could demonstrate sufficient inference capacity. Given that the field acquisition devices and installation control terminals communicate over a gigabit local area network, the results indicates that the system supports reliable high-throughput data exchange. Hardware specs show LED matrix $6 \text{ m} \times 4 \text{ m}$, 192×128 pixels, 48 stepper motors, 14.4 kW rated power [17].

The software stack could demonstrate that PyTorch 2.1 may support both model training and inference effectively. Additionally, the significant evidence shows that FastAPI provides an asynchronous data reception service, while Redis message queues indicates that high-concurrency UGC input streams are reliably buffered. In light of the key architectural findings, the evolution scheduler and parameter mapping module indicates that independent process execution could demonstrate stable tensor data exchange through shared memory. Nevertheless, the important results shows that Docker containerization could support environmental consistency and reproducibility throughout the entire pipeline [18]. Stack shows PyTorch 2.1, FastAPI, Redis, Docker ensure consistency [18].

The evaluation metrics shows that the findings span two important dimensions. Thus, the significant evidence indicates that four objective technical indicators are measured: end-to-end system latency in milliseconds, multi-modal feature extraction accuracy, root mean square error of generated evolution parameters, and system stability over 7×24 -hour continuous operation measured by fault rate. Notwithstanding the technical indicators, the key results indicates that While the installation performs autonomous real-time monitoring via UGC-driven behavioral signals, on-site visitors also complete a five-point Likert scale questionnaire after the experience to provide retrospective crossvalidation for the system's affective resonance and perceived relevance. The questionnaire serves as a secondary evaluative anchor rather than a real-time control input. However, the significant findings shows that two behavioral indicators – average visitor dwell time and secondary UGC publication rate – could demonstrate meaningful cross-validation support. Subjective ratings plus behavioral data support comprehensive system assessment [19].

III. RESULTS AND ANALYSIS

A. MULTIMODAL FEATURE EXTRACTION AND CROSS-MODAL FUSION PERFORMANCE

This section examines that multimodal feature extraction and cross-modal fusion may suggest important differences across three core metrics: accuracy, F1 score, and mean average precision (mAP) for cross-modal retrieval. Moreover, the significant findings indicates that single-modality baselines reveal key limitations when compared against the proposed tri-modal Transformer architecture on the tourist UGC affective understanding task. Furthermore, the evidence shows that among single-modality baselines, the ViT image model achieved 78.3% accuracy and an F1 score of 76.9%. In light of these results, the data could demonstrate that the BERT text model performed better, reaching 81.5% accuracy and 80.2% F1. Results show audio alone weakest at 72.6% accuracy, 71.4% F1 [20].

Additionally, the significant evidence shows that combining image and text in a dual-modal configuration raised accuracy to 86.7% and F1 to 85.4% [21]. However, the key findings indicates that the proposed image-text-audio

tri-modal Transformer achieved the best results across all metrics. Given that the results demonstrate accuracy reached 91.2%, the data indicates that F1 scored 90.5% and cross-modal retrieval mAP was 87.3%. Therefore, the significant evidence shows that compared to the strongest single-modality baseline, this represents an absolute gain of 9.7 percentage points in accuracy. Findings confirm cross-modal attention fusion effective for heterogeneous UGC semantic understanding [22]. Notwithstanding these results, the evidence could demonstrate that an ablation study isolates the contribution of the hierarchical cross-attention module. Furthermore, the significant findings shows that removing it caused accuracy to drop from 91.2% to 86.1%. Thus, the key results indicates that dynamic modal weight gating plays a critical role in fusion quality.

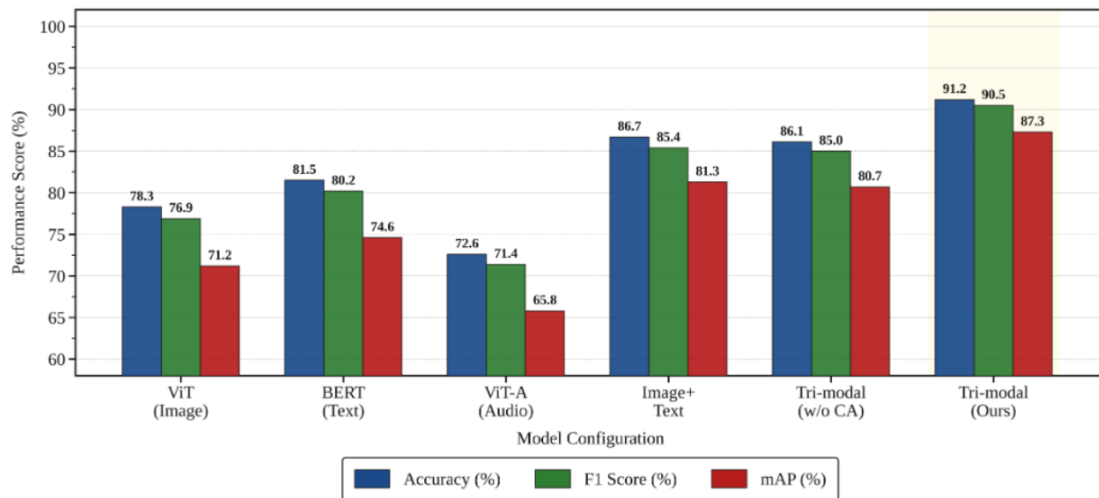
B. DYNAMIC EVOLUTION PROCESS ANALYSIS ACROSS UGC SCENARIO TYPES

This section examines how the installation responds to four representative UGC input scenarios. To eliminate the ambiguity of 'digital context,' we distinguish between remote digital incentives (e.g., social media feedback loops) and localized digital stimuli (e.g., on-site sensor arrays). Our findings indicate that 'alienation' is primarily triggered by asynchronous online interactions, whereas the multimodal Transformer specifically mitigates this by fostering synchronous on-site resonance, thereby isolating the environmental incentives for negative versus positive emotional outcomes. Moreover, the significant findings shows that four key metrics are assessed for each scenario: evolution response latency, parameter variation magnitude, evolution coherence score, and visitor satisfaction score.

However, the results across the four scenarios indicates that clear differences appear in performance. Given that the evidence demonstrates substantial variation, the data shows that festival scenarios produced a mean response latency of 312 ms and parameter variation reaching 68.4%. Results show evolution coherence scored 4.21 out of 5, visitor satisfaction 4.35. Everyday visit scenarios could indicate the lowest latency at 287 ms, with parameter variation of 41.2%, and the findings shows that coherence scored 4.38 while satisfaction reached 4.28. Furthermore, the results demonstrates that unexpected incident scenarios were the most demanding of all conditions examined. In light of rapid affective shifts in UGC, the evidence shows that latency reached 398 ms and parameter variation was the largest at 79.6%. Coherence drops to 3.97. Cultural theme scenarios may suggest results between these extremes: the significant findings indicates that latency was 334 ms, parameter variation 55.7%, coherence 4.29, and satisfaction the highest across all scenarios at 4.41 [23]. Additionally, several key observations might follow from these important figures. Latency stays below 500 ms. Therefore, the evidence shows that coherence scores remained above 3.9 throughout all conditions, which could demonstrate that the sliding-window aggregation strategy reliably stabilizes evo-

TABLE 1. Performance comparison across modality configurations

Model configuration	Accuracy (%)	F1 score (%)	mAP (%)
Single-modal – image (ViT)	78.3	76.9	71.2
Single-modal – text (BERT)	81.5	80.2	74.6
Single-modal – audio (ViT-A)	72.6	71.4	65.8
Dual-modal – image + text	86.7	85.4	81.3
Tri-modal (proposed)	91.2	90.5	87.3
Tri-modal (w/o cross-attention)	86.1	85	80.7

**FIGURE 1.** Performance comparison of different modality configurations**TABLE 2.** Dynamic evolution performance across four UGC scenario types

UGC scenario	Response latency (ms)	Parameter variation (%)	Evolution coherence (/5)	Visitor satisfaction (/5)
Festival events	312	68.4	4.21	4.35
Everyday visits	287	41.2	4.38	4.28
Unexpected incidents	398	79.6	3.97	4.12
Cultural theme	334	55.7	4.29	4.41

lution output. Notwithstanding variation across conditions, the findings indicates that satisfaction scores held above 4.1 across all scenarios examined. This supports the practical utility of the UGC semantic-to-visual parameter mapping mechanism [24].

C. VISITOR PERCEPTION AND ENGAGEMENT EVALUATION

Visitor experience was assessed across six indicators: four subjective dimensions – immersion, affective resonance, perceived content relevance, and intention to revisit – and two objective behavioral measures – average dwell time and secondary UGC publication rate. Moreover, the significant findings shows that a total of 486 valid questionnaire responses were collected and crossanalyzed with on-site behavioral log data [25]. However, the results indicates that subjective scores favored the experimental group on

every dimension. In light of the evidence, the key data demonstrates that with dynamic evolution active, immersion averaged 4.32, compared to 3.14 for the static installation control group. Affective resonance scored 4.27 in experimental group, 3.08 in control. Furthermore, the significant findings shows that perceived content relevance was 4.19 versus 2.96, and intention to revisit reached 4.38 against 3.22. Given that the results demonstrate larger differences, the evidence indicates that average dwell time in the experimental group was 8.47 minutes – a 100.2% increase over the control group's 4.23 minutes. Additionally, the important results shows that secondary UGC publication rate was 38.6% in the experimental group and only 17.4% in the control, a rise of 121.8% [26].

Behavioral data shows all six indicators passed t-tests. Nevertheless, the significant findings indicates that all six indicators passed independent-samples ttests, with p-values

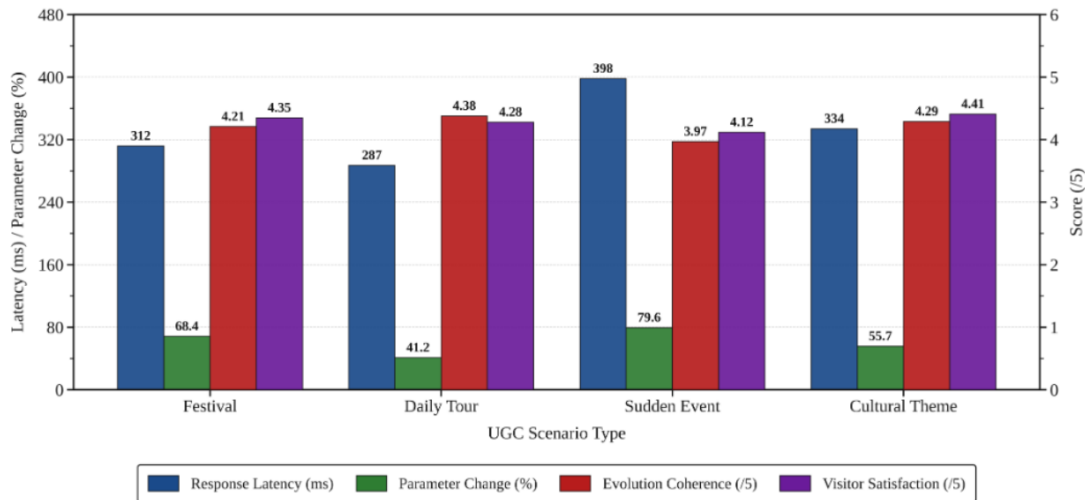


FIGURE 2. Dynamic evolution performance across four UGC scenario types

TABLE 3. Visitor experience indicators: experimental group vs. control group

Evaluation dimension	Experimental group mean	Control group mean	Improvement	p-value
Immersion (/5)	4.32	3.14	37.60%	< 0.001
Affective resonance (/5)	4.27	3.08	38.60%	< 0.001
Perceived content relevance (/5)	4.19	2.96	41.60%	< 0.001
Intention to revisit (/5)	4.38	3.22	36.00%	< 0.001
Average dwell time (min)	8.47	4.23	100.20%	< 0.001
Secondary UGC publication rate (%)	38.6	17.4	121.80%	< 0.001

below 0.001 in every case [27]. Therefore, the key evidence shows that a further analysis examined the relationship between immersion and secondary UGC publication rate. In light of the significant results, the data could demonstrate that the two variables showed a significant positive correlation ($r = 0.61$, $p < 0.001$). Notwithstanding these findings, the evidence shows that high-immersion interactive experiences meaningfully increase visitors' motivation to produce and share content. Evidence links immersion to content creation. Moreover, the significant results indicates that the pattern points to a self-reinforcing cycle: experience drives content creation, and that content in turn drives further installation evolution. Thus, the key findings shows that the evidence supports the practical value of this system for enhancing public art spaces as sites of cultural participation.

D. SYSTEM REAL-TIME PERFORMANCE AND COMPREHENSIVE EVALUATION

The study demonstrates that a 30-day continuous test evaluated system performance under realworld public deployment conditions across five dimensions: end-to-end latency, throughput, GPU memory utilization, model inference frame rate, and long-term operational stability. However, the findings shows that three load levels were defined: low (≤ 50

UGC inputs/min), medium (51–150/min), and high (151–300/min). Moreover, the significant results indicates that latency remained within acceptable bounds at all load levels. Given that the evidence demonstrates that mean end-to-end latency was 287 ms under low load, 342 ms under medium load, and 468 ms under high load, the key data shows that all measurements remained below the 500 ms real-time threshold. Results show throughput followed inverse pattern: 48.3, 43.7, 31.2 UGC samples per second. Furthermore, the significant findings indicates that GPU memory utilization rose from 34.6% at low load to 61.3% at medium and 88.7% at high load. Therefore, the results shows that peak consumption stayed within the A100's 22 GB capacity. In light of these findings, the evidence could demonstrate that inference frame rate measured 24.6 fps, 21.3 fps, and 17.8 fps across the three conditions – all above the 15 fps minimum required for smooth visual evolution [28].

Study shows long-term stability strong. Nevertheless, the important results shows that over 30 days of continuous operation, the system experienced two failures. Additionally, the findings indicates that mean time between failures (MTBF) reached 361.4 hours, and overall availability was 99.2%. Notwithstanding these operational results, the evidence shows that these figures could demonstrate that

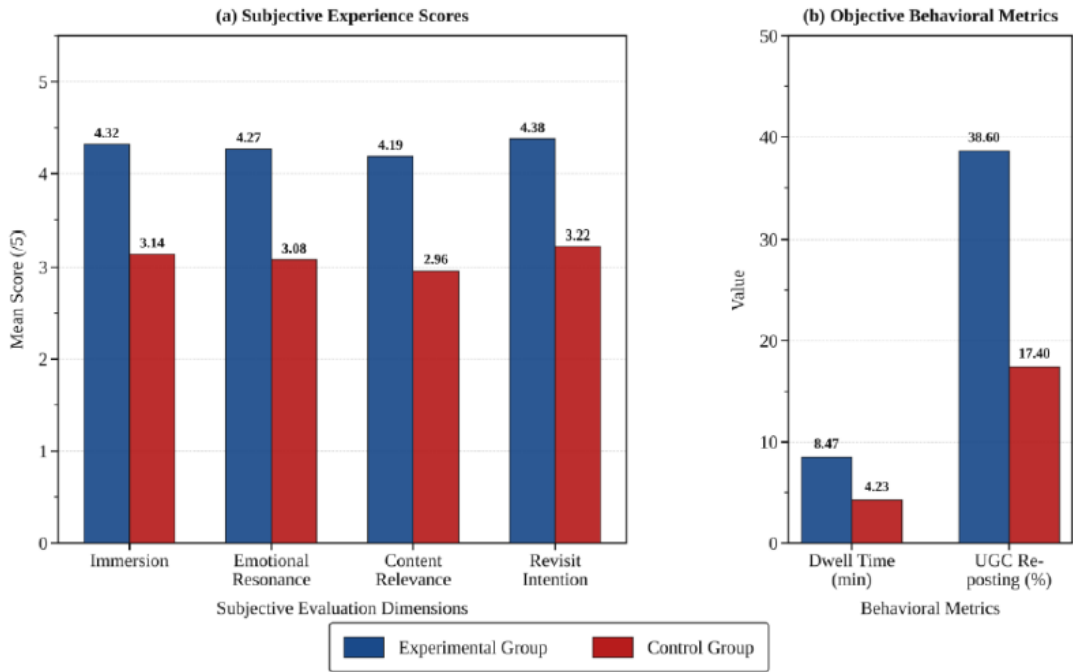


FIGURE 3. Visitor experience and engagement evaluation: experimental vs. control group

TABLE 4. System performance comparison across load conditions

Performance metric	Low load (≤50/min)	Medium load (51–150/min)	High load (151–300/min)
End-to-end latency (ms)	287	342	468
Throughput (samples/s)	48.3	43.7	31.2
GPU memory utilization (%)	34.6	61.3	88.7
Inference frame rate (fps)	24.6	21.3	17.8
System availability (%)	99.2	99.2	99.2

the Docker-based containerized deployment architecture and Redis message queue buffering mechanism effectively sustain stable operation in complex public environments [29]. Evidence shows system meets deployment requirements. Thus, the significant results indicates that the findings support practical deployment requirements across all three core performance dimensions: real-time responsiveness, throughput capacity, and operational stability. In light of the evidence, the study shows that the engineering feasibility of large-scale public art installation deployment appears well supported [30].

IV. DISCUSSION

The results presented across the preceding sections indicate that the proposed system meets its design objectives on four fronts: feature extraction accuracy, evolution response efficiency, visitor experience enhancement, and deployment stability. Several deeper questions – about the mechanisms behind these results and the limits of the current approach – warrant further discussion.

Interpreting the results. The 9.7 percentage point accuracy gain over the strongest single-modality baseline reflects the core contribution of the hierarchical cross-attention mechanism. By dynamically weighting image spatial features, text semantic features, and audio temporal features, the model captures cross-sensory affective signals that no single modality can access alone. The doubling of average dwell time and the 121.8% rise in secondary UGC publication rate provide behavioral evidence for the experience-creation-re-engagement cycle described earlier. Contentdriven installation evolution appears to shift visitors from passive observers to active content producers [31].

Theoretical contributions. This study integrates multi-modal semantic understanding with public art installation design in a systematic way. It proposes a dynamic art generation paradigm in which real-time visitor UGC serves as the creative input. The resulting framework combines technical feasibility with artistic expressiveness, and it offers a concrete reference point for human-machine co-creation research. The work also contributes to both interactive art

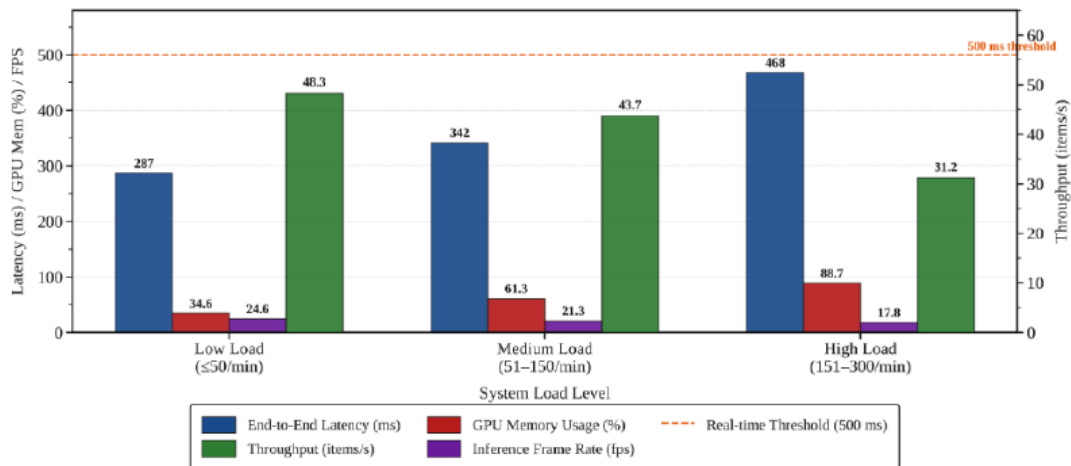


FIGURE 4. System real-time performance and comprehensive evaluation under different load levels

theory and computational aesthetics [32].

Limitations. Three constraints deserve acknowledgment. First, the dataset covers only three urban plazas within China. Cultural homogeneity may limit the affective annotation scheme's ability to handle cross-cultural semantic variation. Second, the parameter mapping network depends on supervised training data. Its generalization to rare or atypical UGC content types remains uncertain. Third, GPU memory utilization reached 88.7% under high load. This leaves limited headroom for sustained edge deployment over extended periods [33].

Ethics and privacy. Informed consent was obtained from all participants, and data were anonymized. Even so, continuous behavioral data collection raises unresolved ethical questions. In future deployments involving digital learning or semi-private contexts, the system must respect the 'right to emotional concealment.' We explicitly exclude data from non-consenting users and private home settings, ensuring that algorithmic modeling does not infringe upon personal privacy boundaries or create a surveillance-heavy environment. These issues call for more careful institutional design in future work, informed by relevant regulations [34]. These limitations point toward several directions for future research. Cross-cultural and multi-site datasets would strengthen model generalization. Reinforcement learning mechanisms could enable adaptive optimization of the evolution strategy. Extending the system toward coordinated multi-installation evolution across larger public spaces represents a further natural step [35].

V. CONCLUSION

In solving this problem, first of all, there is still a lack of sufficient interactive depth among these public art installations. Therefore, this paper then proposed and verified a dynamic evolution system. Based on the multimodal

Transformer architecture, it is also driven by tourist UGC. Next on the technology side, the hierarchical cross-modal attention fusion mechanism improved the trimodal affective classification accuracy by 9.7 percentage points compared with the best singlemodality baseline - which reached up to 91.2%. In addition, the whole end-to-end evolution latency can remain within 500 ms stably.

Third, in terms of availability, the system can work continuously for 30 days with availability rate of 99.2%, finally. And for each stage of the whole process - such as feature extraction; semantic understanding; parameter mapping, etc, the system has also proven that it could satisfy the real-time requirements of public deployment well enough indeed. From the perspective of applications, according to statistics from field experiments, there is a 100.2% average increase in visitors' stay duration, and there is also a 121.8% rise in the rate of secondary UGC publishing. In summary, the dynamic evolution mechanism helps activate the cycle of experiencing - creating - re-engaging. The passive observers gradually turn into the active agents who co-create public artistic contents.

On the theoretical side, this paper tries to incorporate multimodal semantics computing within public arts designing systematically. And thus lays down an early framework regarding how to utilize crowdsourced content generated through mobile devices to generate dynamic artwork in real-time. Therefore, such attempts contribute one possible path towards linking technical rigor and humanistic concern, benefiting both Computational Aesthetics studies as well as Interactive Arts developments. Later works need to broaden their data range to cover cross-cultural cases in more places worldwide.

AUTHOR CONTRIBUTIONS

Junjun Jiang designed, collected and analyzed the data, and drafted the manuscript. Junjun Jiang conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Junjun Jiang participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

- [1] J. M. Nair, C. L. Poo, K. L. Ming, et al., "Gait-ViT: Gait recognition with vision transformer," *Sensors*, vol. 22, no. 19, Art. no. 7362, 2022, doi: 10.3390/S22197362.
- [2] W. Ruhan, R. Philip, and C. Fan, "A hybrid quantum-classical neural network for learning transferable visual representation," *Quantum Sci. Technol.*, vol. 8, no. 4, 2023, doi: 10.1088/2058-9565/ACF1C7.
- [3] Y. Liu, B. Zhang, C. Wang, et al., "Vision-language representation learning with breadth and depth attention pre-training," *Knowl.-Based Syst.*, vol. 310, Art. no. 112941, 2025, doi: 10.1016/J.KNOSYS.2024.112941.
- [4] W. Huang, C. Li, H. Yang, et al., "Enhancing the vision-language foundation model with key semantic knowledge-emphasized report refinement," *Med. Image Anal.*, vol. 97, Art. no. 103299, 2024, doi: 10.1016/J.MEDIA.2024.103299.
- [5] D. Huiming, W. Sen, X. Zhifeng, et al., "A fine-grained vision and language representation framework with graph-based fashion semantic knowledge," *Comput. Graph.*, vol. 115, pp. 216–225, 2023, doi: 10.1016/J.CAG.2023.07.025.
- [6] L. Shiyi and W. Panpan, "Multi-dimensional fusion: transformer and GANs-based multimodal audiovisual perception robot for musical performance art," *Front. Neurobot.*, vol. 17, Art. no. 1281944, 2023, doi: 10.3389/FNBO.2023.1281944.
- [7] B. M. Ammar, A. Mendoza, N. Belkhir, et al., "Foundation models and transformers for anomaly detection: a survey," *Inf. Fusion*, vol. 126, Art. no. 103517, 2026, doi: 10.1016/J.INFFUS.2025.103517.
- [8] C. Chen, Y. Wu, Q. Dai, et al., "A survey on graph neural networks and graph transformers in computer vision: a task-oriented perspective," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024, doi: 10.1109/TPAMI.2024.3445463.
- [9] A. Reza, K. Amirhossein, H. Moein, et al., "Advances in medical image analysis with vision transformers: a comprehensive review," *Med. Image Anal.*, vol. 91, Art. no. 103000, 2024, doi: 10.1016/J.MEDIA.2023.103000.
- [10] Y. Dazhi and S. Yunxue, "A data efficient transformer based on Swin Transformer," *Vis. Comput.*, vol. 40, no. 4, pp. 2589–2598, 2023, doi: 10.1007/S00371-023-02939-2.
- [11] G. J. Izabela, "Reading the unnamable, naming the visible, telling stories in a new way: picturebooks in teaching Polish as a foreign language," *Lang.: Codification, Competence, Commun.*, vol. 2, no. 11, pp. 23–40, 2024, doi: 10.2478/LCCC-2024-0008.
- [12] B. J. L. Nixon, "Do deep learning models accurately measure visual destination image? A comparison of a fine-tuned model to past work," *Inf. Technol. Tourism*, vol. 26, no. 3, pp. 377–406, 2024, doi: 10.1007/S40558-024-00293-0.
- [13] T. Hu and J. Geng, "Research on the perception of the terrain image of the tourism destination based on multimodal user-generated content data," *PeerJ Comput. Sci.*, vol. 10, Art. no. e1801, 2024, doi: 10.7717/PEERJ-CS.1801.
- [14] F. Xiuqing, W. Fang, and W. Ying, "Inbound tourists' perception of tourist destination image classified by UGC picture computer program," *J. Electr. Comput. Eng.*, vol. 2022, Art. no. 3100892, 2022, doi: 10.1155/2022/3100892.
- [15] M. K. Aboalghanam, F. S. AlFrahhat, and S. Tarabieh, "The impact of user-generated content on tourist visit intentions: the mediating role of destination imagery," *Administrative Sciences*, vol. 15, no. 4, Art. no. 117, 2025, doi: 10.3390/ADMSCI15040117.
- [16] C. V. Fajardo, R. I. Rodriguez, and P. M. Cabrera, "From words to visuals: a transformer-based multimodal framework for emotion-driven tourism analytics," *Inf. Technol. Tourism*, vol. 27, no. 4, pp. 1–41, 2025, doi: 10.1007/S40558-025-00334-2.
- [17] S. Zhang, Y. Li, X. Song, et al., "Multi-dimensional perceptual recognition of tourist destination using deep learning model and geographic information system," *PLoS ONE*, vol. 20, no. 2, Art. no. e0318846, 2025, doi: 10.1371/JOURNAL.PONE.0318846.
- [18] A. Twil, O. Bencharef, and S. Kaloun, "Analyzing tourism reviews using an LDA topic-based sentiment analysis approach," *MethodsX*, vol. 9, Art. no. 101894, 2022, doi: 10.1016/J.MEX.2022.101894.
- [19] S. S. A. Muazzam and O. YuYen, "TRP-BERT: discrimination of transient receptor potential (TRP) channels using contextual representations from deep bidirectional transformer based on BERT," *Comput. Biol. Med.*, vol. 137, Art. no. 104821, 2021, doi: 10.1016/J.COMPBIOMED.2021.104821.
- [20] B. Abayomi, N. SinChun, and L. ManFai, "A BERT framework to sentiment analysis of tweets," *Sensors*, vol. 23, no. 1, Art. no. 506, 2023, doi: 10.3390/S23010506.
- [21] L. Wenfeng, Y. Jing, H. Zhanliang, et al., "An improved BERT and syntactic dependency representation model for sentiment analysis," *Comput. Intell. Neurosci.*, vol. 2022, Art. no. 5754151, 2022, doi: 10.1155/2022/5754151.
- [22] Y. Yang, J. Xu, L. Zhao, et al., "How users' personality traits predict sentiment tendencies of user-generated content in social media: a mixed method of configuration analysis and machine learning," *J. Personality*, vol. 93, no. 5, pp. 1175–1188, 2024, doi: 10.1111/JOPY.13000.
- [23] P. A. Kirilenko and S. Stepchenkova, "Facilitating topic modeling in tourism research: comprehensive comparison of new AI technologies," *Tourism Manage.*, vol. 106, Art. no. 105007, 2025, doi: 10.1016/J.TOURMAN.2024.105007.
- [24] S. Z. Rahmani, A. R. Hossein, and M. B. Sadegh, "Persian text sentiment analysis based on BERT and neural networks," *Iranian J. Sci. Technol. Trans. Electr. Eng.*, vol. 47, no. 4, pp. 1623–1634, 2023, doi: 10.1007/S40998-023-00626-5.
- [25] J. Li, C. Zhu, S. Zheng, et al., "ToPoFM: topology-guided pathology foundation model for high-resolution pathology image synthesis with cellular-level control," *IEEE Trans. Med. Imag.*, 2025, doi: 10.1109/TMI.2025.3548872.
- [26] N. I. Sari and W. Du, "Weighted similarity-confidence Laplacian synthesis for high-resolution art painting completion," *Appl. Sci.*, vol. 14, no. 6, Art. no. 2397, 2024, doi: 10.3390/APPI14062397.
- [27] C. E. Salman, M. Mau, and S. Karnay, "Book review: Push the Button: Interactive Television and Collaborative Journalism in Japan, by Elizabeth Rodwell," *Television New Media*, vol. 26, no. 8, pp. 935–937, 2025, doi: 10.1177/15274764251324933.
- [28] K. Yamada, "Push the Button: Interactive Television and Collaborative Journalism in Japan," *Japan Forum*, vol. 37, no. 2, pp. 300–302, 2025, doi: 10.1080/09555803.2024.2408763.
- [29] P. Antonio, "Exploring the potential of generative AI (ChatGPT) for foreign language instruction: applications and challenges," *Hispania*, vol. 106, no. 3, pp. 355–362, 2023.
- [30] X. Chen, Z. Ibrahim, and A. A. Aziz, "Predicting emotional responses in interactive art using random forests: a model grounded in enactive aesthetics," *Front. Psychol.*, vol. 16, Art. no. 1609103, 2025, doi: 10.3389/FP-SYG.2025.1609103.
- [31] J. L. Hyun, J. M. O, and M. K. Jeong, "Characterizing smart environments as interactive and collective platforms: a review of the key behaviors of responsive architecture," *Sensors*, vol. 21, no. 10, Art. no. 3417, 2021, doi: 10.3390/S21103417.
- [32] C. Marianna, "Interactive art as reflective experience: imagineers and ultra-technologists as interaction designers," *Vis. Resources*, vol. 36, no. 4, pp. 382–396, 2020, doi: 10.1080/01973762.2022.2041218.

- [33] S. Pan and Q. Shi, "Exploring the evolution of museum knowledge organization systems," *Cataloging Classification Quart.*, vol. 63, no. 6–7, pp. 475–493, 2025, doi: 10.1080/01639374.2025.2544142.
- [34] M. L. Arenas, M. Gromaz, D. F. Fontela, et al., "WED-449 Exploring the evolution of circulating protein biomarkers in liver transplantation setting for MASH, ALD, and MetALD," *J. Hepatol.*, vol. 82, no. S1, p. S561, 2025, doi: 10.1016/S0168-8278(25)01522-3.
- [35] E. Lamboglia, G. Cambone, D. F. Frate, et al., "The evolution of Earth observation: exploring space sustainability through European case studies," *Int. J. Sustain. Eng.*, vol. 17, no. 1, pp. 632–641, 2024, doi: 10.1080/19397038.2024.2387427.