Cross Modal Fusion Network: Integrating Visual Language Features For Context Aware Visual Communication Design

Ying Zhang

Xinxiang Vocational and Technical College Henan, 453000, China

Corresponding author. E-mail: ying_zhang67@outlook.com

Received: May 09, 2026; Accepted: Aug. 05, 2026

Multimodal information integration is essential for developing context-aware visual communication designs, especially as digital media increasingly relies on flexible and efficient interaction. Traditional approaches often treat text and images as separate elements, limiting contextual coherence and weakening message clarity. This study proposes a cross-modal fusion network that integrates visual and linguistic information to enhance context-aware visual communication. Using the Scene Parse 150 dataset, preprocessing involved tokenization, image resizing, and feature extraction through a convolutional neural network. The model introduces a Self-Attention Based Generative Adversarial-tuned Robustly Optimized BERT Pretraining Approach (SAGA-RoBERTa), where RoBERTa encodes textual descriptions to capture semantic richness and guide a generative adversarial network in producing or refining visual design layouts. A multi-channel fusion module with self-attention further captures intrinsic relationships between modalities, ensuring strong semantic alignment. The discriminator evaluates visual realism and coherence with textual intent. Experimental results demonstrate that SAGA-RoBERTa achieves notable improvements, including an Average IoU of 68.6% and pixel accuracy of 94.5%, outperforming conventional unimodal methods. The findings highlight the potential of cross-modal deep learning frameworks to support more adaptive, emotionally resonant, and semantically precise visual communication across diverse media contexts.

Keywords: Cross-Modal Fusion, Visual-Language Integration, Context-Aware Design, Visual Communication Design, Self-Attention Based Generative Adversarial-tuned Robustly Optimized BERT Pretraining Approach (SAGA-RoBERTa)

http://dx.doi.org/10.6180/jase.202612_35.034

1. Introduction

Visual communication enables individuals to perceive, interpret, and understand information through images, symbols, typography, and layouts, supporting effective message delivery and self-perception [1]. It serves as a powerful communication medium in digital environments by facilitating rapid understanding across different languages and cultures [2]. Language complements visual elements by providing structured meaning, reducing ambiguity, and enhancing semantic clarity through captions, labels, and textual descriptions. The integration of textual and visual information improves accessibility and minimizes misinterpretation of communicative intent [3]. Multimodal communication combines multiple representational modes, such as text, visuals, and audio, to create richer and more effective communication experiences. Each modality contributes unique advantages, including visual presence, linguistic precision, and emotional expression, while redundancy across modalities enhances user engagement and understanding [4]. Context-aware communication design further improves effectiveness by adapting messages according to audience characteristics, cultural factors, environmental conditions, and temporal contexts. Such adaptability

ensures relevance, comprehension, and acceptance in dynamic digital communication environments [5]. Visual and textual elements interact to combine semantic meaning with visual structures, enabling effective communication across applications such as infographics, e-learning, and marketing strategies [6]. Emotional resonance in communication is achieved when visual-linguistic artifacts evoke emotional responses through imagery, composition, color schemes, and narrative expression, thereby enhancing message retention and user engagement [7]. The integration of visual and linguistic modalities strengthens both cognitive understanding and emotional perception, making multimodal communication highly effective in digital media environments. Such multimodal interactions support advertising, education, entertainment, applications, and interactive experiences by improving user engagement and facilitating brand development [8]. Emotional responses are further reinforced through the complementary roles of visual composition and language-based storytelling [9]. However, continuous contextual adaptation introduces challenges related to data acquisition, computational cost, privacy concerns, and potential biases arising from incomplete contextual information. Additionally, implementing adaptive communication frameworks requires substantial infrastructure, interdisciplinary expertise, and resource investment, which may limit adoption and integration within existing systems [10]. Existing visual-language integration methods often face challenges in maintaining semantic consistency and contextual relevance across modalities. To address these limitations, the proposed CMFN combines self-attention fusion, GAN-guided layout generation, and RoBERTa semantic encoding to achieve improved cross-modal alignment, contextual adaptability, and visual communication performance. The research aims to develop and validate a Cross-Modal Fusion Network (CMFN) that integrates visual and textual features to generate context-aware visual communication designs. The proposed SAGA-RoBERTa framework enhances semantic alignment, contextual relevance, and adaptive visual layout generation through multimodal learning.

- Introduces a self-attention and adversarially tuned RoBERTa model for improved semantic representation learning and textual-visual coherence.
- Implements multi-channel self-attention and decision-level fusion to capture cross-modal relationships and enhance robustness and adaptability.
- Achieves superior semantic alignment, contextual relevance, and user engagement compared to conventional approaches, supporting adaptive visual communication design.

Section 2 reviews related work on context-aware visual communication design. Section 3 presents the proposed CMFN with SAGA-RoBERTa methodology. Section 4 evaluates semantic alignment and performance metrics, while Section 5 concludes the study and outlines future directions.

2. Related work

The SLRGAN model improved context-aware continuous sign language recognition using GAN-based learning by incorporating contextual information beyond conventional spatiotemporal features [11]. The DLVPR-MCATS framework combined ResNet, EfficientNet, and MobileNetV2 features to enhance visual place recognition and multimedia communication performance in autonomous transportation systems [12]. A context-aware multimodal fusion framework for Bengali sarcasm and humor detection integrated MobileNetV2 visual features with BiLSTM and GRU-based textual representations, achieving high classification accuracy [13]. The BICA deep network enhanced watermark robustness and image quality, maintaining nearly 95% extraction accuracy under various noise attacks. Recent transformer-based multimodal and vision-language foundation models have advanced cross-modal understanding through large-scale pretraining and attention mechanisms, although their application to context-aware visual communication remains limited [14]. Liu and Yang [15] proposed a meta-learning framework for multi-scene student posture recognition and performance assessment, demonstrating the effectiveness of adaptive representation learning across diverse environments. Similar to the proposed SAGA-RoBERTa framework, the study highlights the importance of contextual feature modeling and robust learning strategies for improving performance in complex real-world scenarios.

2.1. Research gap

Existing studies in multimodal and visual communication design have introduced innovative frameworks, yet several research gaps persist. SLRGAN focused on spatiotemporal elements in sign language recognition but neglected textual or contextual cues, limiting semantic richness [11]. DLVPR-MCATS improved multimedia communication with hybrid feature fusion but relied heavily on optimization without addressing semantic alignment [12]. Multimodal sarcasm detection in Bengali demonstrated strong performance but remained domain-specific and language-limited [13]. Sim-

ilarly, BICA enhanced watermark robustness but concentrated on visual fidelity without broader contextual integration [14]. Current studies face challenges in preserving semantic similarity, adapting contextual representations, stabilizing adversarial multimodal training, and generating emotionally engaging visual communication outputs.

3. Materials and methods

The proposed methodology integrates multimodal textual and visual data through preprocessing, image resizing, and contextual data collection to enrich representation learning. CNN-based feature extraction is employed for dimensionality reduction, while the proposed SAGA-RoBERTa framework combines self-attention-driven multimodal fusion, adversarial learning, and RoBERTa semantic embeddings to perform context-aware visual layout generation and refinement. This architecture enhances semantic consistency, visual realism, and contextual relevance across multimodal inputs. Fig. 1 illustrates the overall architecture of the proposed framework.

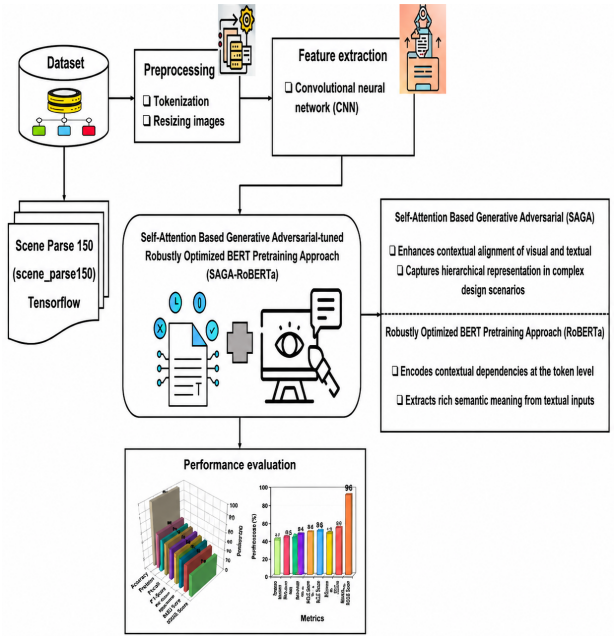

Fig. 1. Overall architecture of the proposed model.

3.1. Dataset

Scene Parse 150 (scene_parse150) Tensorflow dataset is the process of segmenting and parsing an image into different image regions according to semantic categories such as sky, road, person, and bed from open source in Kaggle (https://www.kaggle.com/datasets/naresh/scene-parse-150-tf?select=scene_parse150). The MIT Scene Parsing Benchmark (SceneParse150) is a standardized platform for training and evaluating scene parsing algorithms. Fig. 2 shows the sample images of the dataset. SceneParse150 was selected because it contains diverse scene categories and rich semantic annotations that facilitate contextual representation learning and multimodal feature alignment. Its comprehensive segmentation labels provide a suitable benchmark for evaluating context-aware visual communication and semantic consistency.

Fig. 2. Overall example images of the dataset.

3.2. Preprocessing using tokenization and Resizing images

Tokenization converts textual descriptions into words or subwords for effective semantic encoding and alignment with visual elements. Image resizing standardizes image dimensions for consistent CNN-based feature extraction while reducing computational complexity. Together, these preprocessing steps enhance contextual understanding and support coherent context-aware visual communication design. Fig. 3 illustrates the image preprocessing workflow adopted in the proposed framework. The input images from the SceneParse150 dataset are resized to a uniform resolution and normalized to ensure consistency across samples. This preprocessing step reduces computational complexity, preserves essential semantic content, and facilitates effective synchronization between visual features and corresponding textual representations during multimodal learning. Fig. 3 shows the preprocessed images for the original vs. the resized.

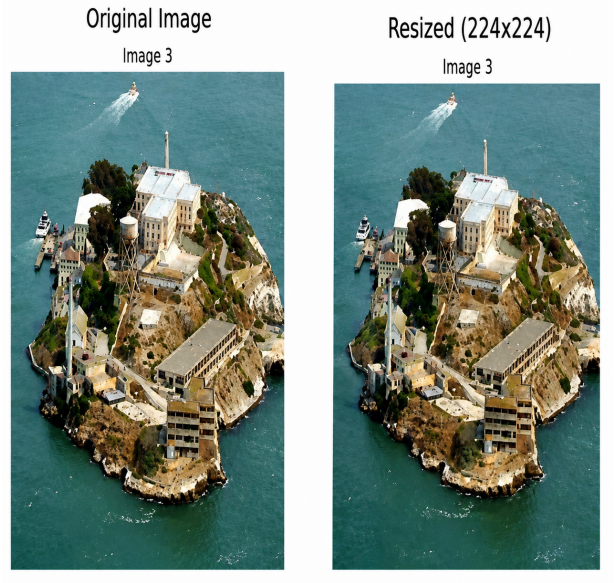

Fig. 3. Overall example images of the dataset.

3.3. Preprocessed images for original vs. resized

The CNN module extracts local spatial patterns while suppressing noise and generating hierarchical feature representations that capture contextual information. A Transformer encoder further enhances these features by focusing on important regions, and the resulting attention-enhanced embeddings support effective multimodal fusion, semantic alignment, and context-aware visual communication design. The input sequence is represented as a matrix, where each row corresponds to the embedding vector of a word or character. An input data matrix J (such as an image or spatial signal) is processed through a set of learnable filters E, resulting in a collection of feature maps X (Eq. (1)).

$$\mathrm{E}(j,i,l) = \sum_{n,m} J(j+n, i+m) X(n,m,l) + a_l \qquad (1)$$

3.4. Self-Attention Based Generative Adversarial-tuned Robustly Optimized BERT Pretraining Approach (SAGA-RoBERTa) for context-aware visual communication design

The proposed SAGA-RoBERTa framework combines RoBERTa contextual embeddings, self-attention learning, GAN-based adversarial optimization, and robust training strategies to support context-aware visual communication design. Self-attention enhances textual-visual alignment, while discriminator–generator interaction improves visual realism and semantic consistency, resulting in stable multimodal learning and effective semantic alignment preservation. Fig. 4 shows the feature maps extracted from different layers of the CNN, illustrating how the network progressively learns low-level patterns, edges, textures, and higher-level discriminative features from the input image through convolution and pooling operations.

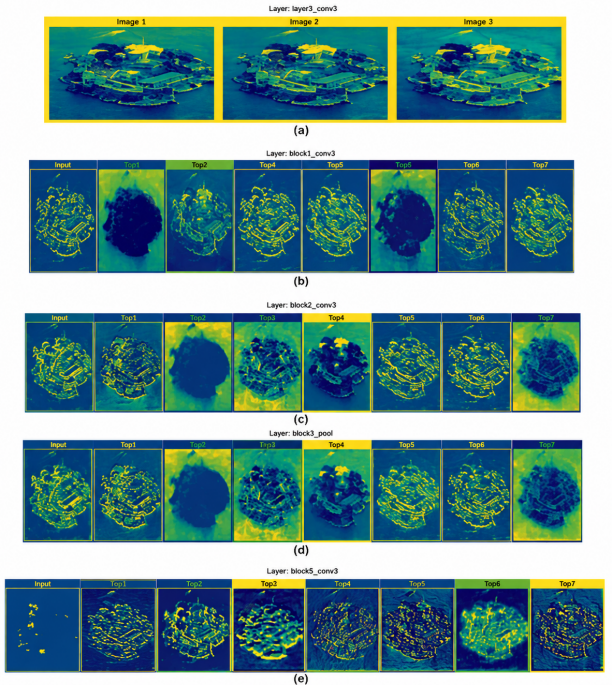

Fig. 4. Overall extracted the dataset images (a) input_layer_1, (b) block1_conv1, (c) block1_conv2, (d) block1_pool, (e) block2_conv1.

3.4.1. *Self-Attention Based Generative Adversarial Network (SAGA)*

Beyond image realism, GANs can be modified to include visual-language aspects, in which textual semantics lead the generator in producing designs or images that are consistent with linguistic context, while the discriminator assesses both visual authenticity and semantic coherence. This goal permits the creation of outputs that are both visually appealing and contextually appropriate, promoting improved communication in multimodal applications (**??**).

$$\min_H \max_C U(C,H) = \mathbb{E}_{w\sim p_{\mathrm{data}}(w)}[\log C(w)] + \mathbb{E}_{y\sim p_y(y)}[\log(1 - C(H(y)))] \qquad (2)$$

The minimax optimization method, in which the generator $\min_H$ attempts to minimize the loss and the discriminator $\min_C$ aims to maximize. Here, $\mathbb{E}_{w\sim o_{\mathrm{data}}(w)}$ denotes actual data samples taken in the true data distribution, while $\mathbb{E}_{y\sim o_y(y)}$ depicts latent variables taken from a preceding noise distribution. The term $\log C(w)$ encourages the discriminator to assign high probability to real samples,

whereas $1 - C(H(y))$, the discriminator properly detects created samples as fraudulent, and the generator is penalized. This adversarial training pushes C to produce realistic outputs, while H continuously improves its discriminative capability. Selfattention enhances multimodal learning by selectively focusing on informative features and modeling global relationships between visual and textual representations. This mechanism improves semantic alignment and contextual coherence by capturing cross-modal dependencies beyond the limitations of conventional convolutional operations. This selfattention process is mathematically stated in Eq. (3).

$$\text{Attention}(R, L, U) = \text{softmax}\left(\frac{RL^S}{\sqrt{c_l}}\right) U \tag{3}$$

Where R, L, and U depict the many linear projections of the same data, which is represented by the query, key, and value matrices that are obtained from the input features. The operation RL^S computes similarity scores between queries and keys, with S indicating the transpose to align dimensions for dot-product calculation. The scaling factor $\sqrt{c_l}$, where c_l is the dimension of the key vectors, normalizes these scores to prevent excessively large values. Applying the softmax function transforms the scores into probability weights, which are then used to aggregate the values of U, yielding a contextually weighted representation that emphasizes the most relevant features.

3.4.2. *Robustly Optimized BERT Pretraining Approach (RoBERTa)*

To enhance downstream job performance, the RoBERTa makes numerous significant changes to the original BERT framework. Stronger contextual awareness is made possible by RoBERTa's reliance on full-sentence training rather than the Next Sentence Prediction (NSP) objective, which BERT uses. This corruption process is applied during training to encourage the model to recover the original image input, thereby improving its robustness and denoising capability, as follows in Eq. (4).

$$K_{\text{total}}(\theta) = \lambda K_{\text{denoise}}(\theta) + \sum_j W_{\text{task}_j}(\theta) \tag{4}$$

The overall loss function $K_{\text{total}}(\theta)$ combines the denoising objective and multiple task-specific objectives in visual communication design. The coefficient λ is a tunable hyperparameter that strengths the role of the denoising loss $K_{denoise}(\theta)$ against the individual downstream task losses $W_{task\ j}(\theta)$.Each $W_{task\ j}$ quantifies performance on a specific learning task, ensuring the model jointly learns robust representations while maintaining task adaptability. The RoBERTabase encoder utilizes 12 attention heads to simultaneously capture diverse semantic patterns and contextual relationships from multimodal inputs. This multi-head attention mechanism enhances feature learning by modeling complementary dependencies between textual and visual representations, leading to improved semantic alignment and contextual understanding. The model uses RoBERTabase with 12 transformer layers, 768 hidden dimensions, and 12 attention heads for robust textual feature encoding.

4. Result and discussion

The generated layouts incorporate essential visual communication principles such as hierarchy, contrast, and compositional balance, enhancing message clarity and user engagement. For model evaluation, the dataset was partitioned into 80% training and 20% testing samples, and training was performed using the Adam optimizer with a learning rate of 0.0001, batch size of 32, and 100 epochs on an NVIDIA CUDA-enabled GPU. Convergence was achieved when the validation loss remained stable without significant improvement over consecutive training iterations. The system configuration used for training and evaluation. The experiments were conducted using an NVIDIA CUDA-enabled GPU, 16 GB RAM, Python 3.8, and the Windows 10 operating system.

Fig. 5 shows the ROC curve demonstrating the strong discriminative capability of the suggested model, achieving a high area under the curve, and reflecting superior separation between positive and negative prediction classes. The ROC curve demonstrates strong discriminative capability of the proposed model, achieving an AUC of 0.969, indicating excellent separation between positive and negative classes.

Fig. 6 depicts the precision-recall curve, which indicates a balance of precision and recall besides underlining the resilience of the proposed strategy in dealing with class imbalance and enhancing positive case detection. The decision threshold was chosen near the optimal precision–recall balance point, providing reliable classification performance across varying decision boundaries and class distributions.

The comparative analysis evaluates the SAGA-RoBERTa method against existing methods such as objectless region enhancement network (OENet) [16], Segmentation Network (SegNet) [16], Cascade-SegNet [16], Fully Convolutional Network with 8-pixel stride (FCN8 s) [16], Dilated Convolutional Network (DilatedNet) [16], Deep CNN for Semantic Segmentation (Version 2) (DeepLabv2) [16], Cascade Dilated Convolutional Network (Cascade-DilatedNet) [16], and Bascline [16] for context-aware visual communication design, highlighting its improved feature extrac-

Table 1. Numerical outcomes of the existing and proposed methods.

Methods	Average IoU (%)	Pixel accuracy (%)
SegNet [16]	21.5	70.0
Cascade-SegNet [16]	27.4	71.7
FCN8 s [16]	29.4	71.2
DilatedNet [16]	32.2	73.5
DeepLabv2 [16]	34.2	75.2
Cascade-DilatedNet [16]	34.8	74.4
Bascline [16]	30.8	73.0
OENet [16]	38.3	77.8
SAGA-RoBERTa	68.6	94.5

Table 2. Evaluation Metrics of the Proposed SAGA-RoBERTa Framework.

Method	Precision (%)	Recall (%)	F1-Score (%)	Semantic Alignment Score (%)
SegNet	72.4	69.8	71.1	65.2
Cascade-SegNet	74.6	72.1	73.3	67.8
FCN8s	75.9	73.5	74.7	69.1
DilatedNet	77.3	75.2	76.2	71.4
DeepLabv2	79.1	77.0	78.0	73.5
Cascade-DilatedNet	80.2	78.4	79.3	74.6
Baseline	78.0	76.2	77.1	72.3
OENet	83.6	81.8	82.7	79.5
SAGA-RoBERTa (Proposed)	96.1	94.4	95.2	92.8

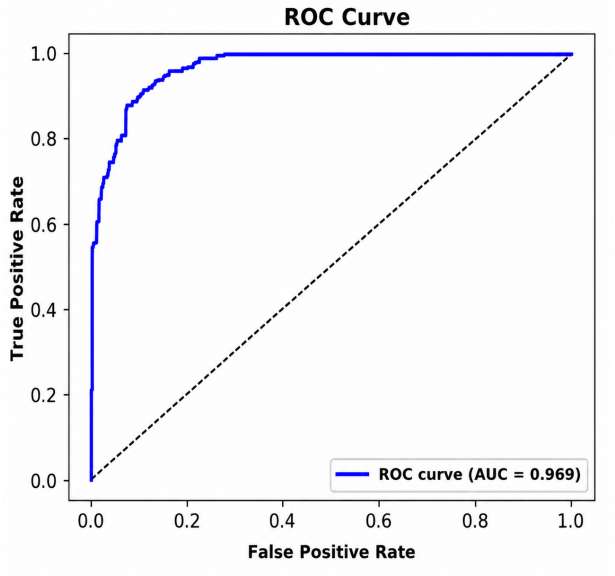

Fig. 5. Receiver Operating Characteristic curve explaining classification model performance.

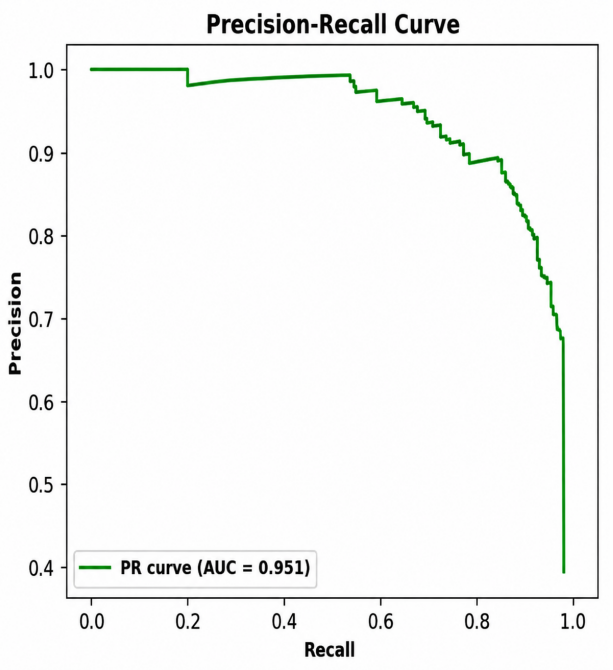

Fig. 6. Receiver Operating Characteristic curve explaining classification model performance.

tion, classification accuracy, and robustness, resulting in enhanced detection precision and reliability over existing approaches. Table 1 displays the numerical results of the existing and proposed methods.

Average Intersection over Union (IoU): Measures overlap between predicted and actual design regions, ensuring contextual alignment of text-image elements. Higher IoU indicates improved semantic accuracy in multimodal communication layouts.

Pixel accuracy: Represents the proportion of correctly classified pixels in visual design tasks. Ensures text and imagery elements are contextually preserved. Higher accu-

racy supports effective, coherent, and context-aware communication outcomes. Compared with OENet, the proposed SAGA-RoBERTa achieves a substantial improvement in Average IoU, increasing from 38.3% to 68.6%, which represents a 79.1% enhancement. Furthermore, Pixel Accuracy improves from 77.8% to 94.5%, demonstrating a 21.5% gain and highlighting the effectiveness of the proposed multimodal fusion framework.

As shown in Table 2, the proposed SAGA-RoBERTa framework achieved the highest Precision (96.1%), Recall (94.4%), and F1-score (95.2%), demonstrating its ability to accurately identify contextually relevant visual-textual relationships while minimizing false detections. Furthermore, the Semantic Alignment Score reached 92.8%, indicating strong consistency between textual semantics and generated visual layouts. These results further validate the effectiveness of the self-attention fusion mechanism, RoBERTa-based semantic encoding, and GAN-guided optimization in enhancing contextual understanding and multimodal communication performance. The superior scores across all evaluation metrics confirm the robustness and reliability of the proposed framework for context-aware visual communication design.

The proposed SAGA-RoBERTa framework effectively handles visually complex scenes through cross-modal semantic fusion and self-attention learning, preserving contextual consistency and reducing ambiguity in diverse visual environments. Qualitative analysis demonstrates that the generated layouts maintain strong semantic alignment between textual descriptions and visual elements, thereby enhancing message comprehension and communication effectiveness. The superior performance of the framework is attributed to the integration of RoBERTa-based semantic encoding, self-attention-driven multimodal fusion, and GAN-based optimization, which collectively improve contextual understanding and visual-language alignment. Furthermore, cross-modal semantic guidance enables more coherent and contextually relevant layout generation. Comparative evaluation against established segmentation methods, including SegNet, FCN8s, DilatedNet, Cascade-DilatedNet, DeepLabv2, and OENet, validates the effectiveness of the proposed approach, while future studies extend benchmarking to transformer-based multimodal architectures and vision-language foundation models.

Ablation study

An ablation analysis was conducted by individually removing the self-attention module, GAN optimization, RoBERTa encoder, and fusion mechanism. Results indicate that each component contributes positively to performance, with the complete architecture achieving the highest semantic alignment and segmentation accuracy.

5. Conclusion

This research highlights the significance of integrating multimodal features to advance context-aware visual communication design. The proposed CMFN with the SAGA-RoBERTa framework successfully aligns textual semantics with visual representations, enabling more adaptive and meaningful design generation. Experimental findings demonstrate superior performance, with improvements in average IoU (68.6%), and pixel accuracy (94.5%), showing the model's robustness in semantic alignment, contextual relevance, and user engagement compared to unimodal approaches. Despite these advancements, limitations remain, such as color inconsistencies in generated outputs and layout design challenges in complex scenarios, which affect the overall design quality. Future research can address these gaps by incorporating color consistency mechanisms, advanced generative modules, and reinforcement learning-based adaptive design strategies, thereby enhancing flexibility, interpretability, and creative control in multimodal communication systems. Strong semantic alignment between textual and visual elements improves message comprehension and contextual relevance in visual communication design. Future work will investigate color harmony analysis, scalable model architectures, computational optimization, real-time deployment, domain adaptation, and cross-domain evaluation to enhance communication effectiveness and generalization. Further research will address adversarial training instability, semantic drift, modality imbalance, contextual ambiguity, computational complexity, and memory consumption, while exploring adaptive real-time systems and comparisons with state-of-the-art multimodal and vision-language models to validate the scalability and robustness of the proposed SAGA-RoBERTa framework.

Declarations

Declaration of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

No dataset was generated or analyzed in this study

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this research.

Funding Statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author Contributions

Ying Zhang conceptualized the research idea, designed the methodology, and supervised the study. The author also conducted the experimental analysis, interpreted the results, and prepared the manuscript.

Ethical Approval

This research does not involve human participants, animals, or biological materials. Therefore, ethical approval was not required.

Consent to Participate

Not applicable. This study does not involve human participants.

Consent for Publication

The author has read and approved the final version of the manuscript and consents to its publication.

Competing Interests

The author declares that there are no competing interests related to this study.

Code Availability

The code and implementation details supporting the findings of this study are available from the corresponding author upon reasonable request for academic and research purposes.

Acknowledgement

The author would like to thank Xinxiang Vocational and Technical College, Henan, China, for providing academic support and research facilities that made this study possible.

References

[1] J. Zhu, (2025) *"Visual contextual perception and user emotional feedback in visual communication design"* **BMC Psychology *13***: 1–13. DOI: 10.1186/s40359-025-02615-1.

[2] C. Zhang, W. Zeng, and L. Liu, (2021) *"UrbanVR: An immersive analytics system for context-aware urban design"* **Computers & Graphics *99***: 128–138. DOI: 10.1016/j.cag.2021.07.006.

[3] H. Elfaik, (2021) *"Combining context-aware embeddings and an attentional deep learning model for Arabic affect analysis on Twitter"* **IEEE Access *9***: 111214–111230. DOI: 10.1109/ACCESS.2021.3102087.

[4] A. Manolova, K. Tonchev, V. Poulkov, S. Dixir, and P. Lindgren, (2021) *"Context-aware holographic communication based on semantic knowledge extraction"* **Wireless Personal Communications *120***: 2307–2319. DOI: 10.1007/s11277-021-08560-7.

[5] L. Lu and L. Huang, (2022) *"Exploration and application of graphic design language based on artificial intelligence visual communication"* **Wireless Communications and Mobile Computing *2022***: 9907303. DOI: 10.1155/2022/9907303.

[6] Z. Xie, B. Zhou, X. Cheng, E. Schoenfeld, and F. Ye, (2022) *"Passive and context-aware in-home vital signs monitoring using co-located UWB-depth sensor fusion"* **ACM Transactions on Computing for Healthcare *3***: 1–31. DOI: 10.1145/3549941.

[7] A. Omolaja, A. Otebolaku, and A. Alfoudi, (2022) *"Context-aware complex human activity recognition using hybrid deep learning models"* **Applied Sciences *12***: 9305. DOI: 10.3390/app12189305.

[8] A. Montanha, A. M. Oprescu, and M. Romero-Ternero, (2022) *"A context-aware artificial intelligence-based system to support street crossings for pedestrians with visual impairments"* **Applied Artificial Intelligence *36***: 2062818. DOI: 10.1080/08839514.2022.2062818.

[9] Z. Ren, L. He, and J. Lu, (2023) *"Context-aware edge-enhanced GAN for remote sensing image super-resolution"* **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing *17***: 1363–1376. DOI: 10.1109/JSTARS.2023.3333271.

[10] X. Chen, J. Li, T. Gao, Y. Piao, H. Ji, B. Yang, and W. Xu, (2024) *"Dynamic context-aware pyramid network for infrared small target detection"* **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**: DOI: 10.1109/JSTARS.2024.3434330.

[11] A. Papastratis, K. Dimitropoulos, and P. Daras, (2021) *"Continuous sign language recognition through a context-aware generative adversarial network"* **Sensors *21***: 2437. DOI: 10.3390/s21072437.

[12] T. Abirami, A. K. Dutta, and S. Alsubai, (2024) *"Deep learning-powered visual place recognition for enhanced mobile multimedia communication in autonomous transport systems"* **Alexandria Engineering Journal *109***: 950–962. DOI: 10.1016/j.aej.2024.09.060.

[13] M. Rahman, M. A. M. Provath, K. Deb, P. K. Dhar, and T. Shimamura, (2025) *"CAMFusion: Context-aware multi-modal fusion framework for detecting sarcasm and humor integrating video and textual cues"* **IEEE Access**: DOI: 10.1109/ACCESS.2025.3535694.

[14] A. Yin and K. Yin, (2025) *"Robust image watermarking using bidirectional-interactive and context-aware networks"* **IEEE Transactions on Circuits and Systems for Video Technology**: DOI: 10.1109/TCSVT.2025.3543969.

[15] L. Liu and J. Yang, (2025) *"A meta-learning-based multi-scene student posture detection method for enhancing learning motivation and engagement in smart classrooms"* **Journal of Applied Science and Engineering *29***(3): 707–724. DOI: 10.6180/jase.202603_29(3).0021.

[16] L. Lu, (2020) *"Design of visual communication based on deep learning approaches"* **Soft Computing *24***: 7861–7872. DOI: 10.1007/s00500-019-03954-z.