Evaluation of cross-cultural communication effectiveness of advertising creative based on adversarial sample robustness test

Yanan Lin

School of Business,
Zhengzhou Shengda University,
Zhengzhou, 451191, China
Email: YananLinn@163.com

Abstract: With the growing use of deep learning in advertising content generation, model robustness in cross-cultural scenarios has become increasingly critical. Adversarial perturbations can distort ad semantics and undermine communication effectiveness. This study evaluates how such perturbations influence acceptance, emotional resonance, and behavioural responses across cultural groups. We propose an integrated framework combining adversarial robustness testing and cultural adaptation assessment, construct a cross-cultural communication platform, and design adversarial attacks for image and text ads based on real data. Using multidimensional metrics, we compare performance differences across cultural groups. Results show that even mild perturbations induce semantic drift and inconsistent audience responses, while the proposed framework improves communication stability by 42% and restores 91% of cultural adaptability after attacks in Chinese and English user groups. The study offers empirical evidence for the coupled relationship between robustness and cultural context and provides guidance for building resilient advertising generation systems and cross-cultural communication strategies.

Keywords: adversarial samples; robustness test; advertising communications; cross-cultural communication; deep learning; content generation.

Reference to this paper should be made as follows: Lin, Y. (2026) ‘Evaluation of cross-cultural communication effectiveness of advertising creative based on adversarial sample robustness test’, *Int. J. Information and Communication Technology*, Vol. 27, No. 1, pp.20–38.

Biographical notes: Yanan Lin graduated with a Bachelor’s degree from Henan Normal University in 2006 and Master’s in Communication from Zhengzhou University in 2011. She studied at Lyceum of the Philippines University from 2019 to 2023 and obtained her PhD in Business Administration. Currently, she works in Zhengzhou Shengda University. She has published ten papers and a book. Her research interests are included advertising and communication. She has guided students to win first prizes in national competitions.

1 Introduction

Today, global advertising increasingly relies on artificial intelligence, and the security and reliability of advertising content are particularly critical. Assuming that a malicious attacker only makes minor modifications to the ad creative, it may cause the ad to fail to spread normally in a specific region. This "cross-cultural communication failure" will not only affect the brand image, but may also cause huge economic losses. In order to counter this risk, it becomes particularly important to study the communication effect of advertising creativity in different cultural contexts and its robustness to potential attacks. With the wide application of deep learning technology in the fields of computational advertising, content generation, and intelligent recommendation, automated advertising creation system has gradually become an important technical foundation for cross-platform marketing (Sayyad et al., 2025). Relying on large-scale pre-trained models, these systems can quickly generate multimodal advertising materials, such as graphics, text, and videos, according to user profiles, product characteristics, and platform preferences, thereby significantly improving content production efficiency and delivery accuracy (Sahbi et al., 2025). However, under the background of global advertising, advertising content needs to face user groups from different cultural backgrounds, language systems and value systems, and its communication effect no longer only depends on the aesthetic and logical structure of the content itself, but also is deeply influenced by cross-cultural factors such as semantic understanding, cultural identity and symbolic habits (Hoffmann et al., 2014). At the same time, recent research on adversarial samples has revealed the high vulnerability of deep learning models to small disturbances. Even input disturbances that are difficult for humans to detect may cause the model to make serious semantic deviations in the generation or classification process output. Suppose this vulnerability occurs during the process of advertising generation and delivery. In that case, it may not only weaken the communication effect but also lead to problems such as cultural misunderstandings and brand reputation crises. Therefore, it is of great theoretical value and practical significance to study the robustness of the advertising generation system in cross-cultural communication scenarios, especially its performance under conditions of resisting sample attacks (Wang et al., 2024c; Gao et al., 2024). In this article, adversarial samples refer to data that may cause the artificial intelligence model to make wrong judgments after minor modifications to the input data; robustness refers to the ability of the model to maintain correct prediction and stable performance in the face of these small disturbances. These two concepts form the core basis for evaluating the stability and security of systems when we study cross-cultural advertising communication.

Adversarial sample attacks were initially widely studied in image classification tasks and have gradually expanded to text, speech, and multi-modal tasks, posing severe challenges to the security of deep models (Wang et al., 2024a). Numerous studies have recently focused on enhancing the adversarial robustness of models. Common methods include adversarial training, model regularisation, and input disturbance detection, among others. At the same time, research in the field of advertising communication primarily focuses on content optimisation, emotion modelling, user behaviour prediction, and other related issues; however, research on the robustness of advertising materials remains underdeveloped (Lv et al., 2026). Especially in a cross-cultural communication environment, different languages and cultures exhibit significant differences in the interpretation of content semantics, which can lead to confrontational disturbances appearing as meaningless changes in one culture but causing misunderstanding or even offence in another culture (Dong, 2025). For example, an English advertisement can still retain its original meaning in the English context after generating adversarial text through slight word modifications, but it may completely lose its contextual coherence when machine-translated into a certain language. This kind of phenomenon is also reflected in image advertisements, such as changes in character posture, colour contrast, and symbols, which may be interpreted completely differently by the audience due to cultural differences (Belanche et al., 2025).

In addition, the rapid development of generative artificial intelligence in recent years has further automated and modelled advertising creation, which not only improves efficiency but also intensifies the trend of 'black boxing' the system. On the one hand, the results generated by the model are difficult to interpret, and it is difficult to identify potential robustness risks. On the other hand, current generative systems generally lack modelling mechanisms for 'cultural sensitivity', making it easier to trigger understanding biases in different cultural environments (Kim et al., 2022). Therefore, it is urgent to build a new framework that integrates adversarial robustness testing and cultural adaptation evaluation to comprehensively examine the stability and communication effect of advertising generation systems in multicultural scenarios.

However, the existing researches on adversarial samples mainly focus on the security and robustness of general models, and often ignore the communication characteristics and cultural sensitivity in different cultural backgrounds. Different languages, customs, and aesthetic preferences may lead to very different reception effects of the same content when advertising creative is spread across cultural environments. If these differences are analysed by combining adversarial sample generation technology, we can systematically evaluate the robustness of advertising creativity in multicultural groups. Therefore, in order to fill the above research gaps, this paper proposes a cross-cultural communication robustness evaluation framework for advertising generation system, which aims to systematically test and compare the communication performance of advertising materials in multicultural groups by constructing a cross-language and cross-cultural communication experimental platform and combining adversarial sample generation technology. In the experiment, we selected representative multilingual advertising samples (covering English, Chinese, Spanish, etc.), and designed anti-perturbation methods based on various generation mechanisms (such as image smoothing perturbation, text semantic attack), and then analysed their changes in acceptance, emotional resonance intensity, and comprehension consistency in different cultural groups. At the same time, we introduce a set of multi-dimensional communication effect indicators, including: content acceptance score (Wang et al., 2024b), semantic consistency (Pasqualette and Kulke, 2024), emotional shift (Betancur Marquez and de Klerk, 2025) and conversion intent drift (Zhang et al., 2024b), which is used to quantify the potential impact of perturbation on communication effect. The purpose of this study is to explore whether the small disturbance of adversarial sample generation will have different effects on the advertising communication effect of audiences with different cultural backgrounds.

The main contributions of this paper are as follows:

1 Propose the first cross-cultural adversarial robustness evaluation framework for an advertising generation system, and reveal the potential semantic shift and understanding risks that may be caused by adversarial disturbance in cross-cultural communication.

2 Construct a multilingual and multicultural advertising communication data set, and carry out empirical tests based on real communication feedback data to provide real scene support for robustness evaluation.

3 Introduce a set of systematic communication effect evaluation indicators, taking into account the three dimensions of semantics, emotion and behavioural intention, and improving the interpretability and multidimensionality of evaluation.

4 Based on the experimental results, put forward targeted improvement suggestions, covering directions such as generative strategy optimisation, cultural sensitivity modelling and multi-modal collaborative robust training.

The purpose of this paper is to provide theoretical support and methodological innovation for the security design and communication of an advertising intelligent generation system in the context of globalisation, and to promote the development of a deep model from ‘available’ to ‘reliable’ and from ‘efficient’ to ‘credible’. The conclusions of this study can provide practical guidance for advertising security audit and localisation strategies of ‘multinational enterprises’, which can improve the importance of the article.

2 Theoretical knowledge related to the evaluation of cross-cultural communication effect of advertising creative against sample robustness test

2.1 Basic theory of adversarial samples

Adversarial Examples refer to the input samples that add carefully designed tiny perturbations to the input samples, causing the deep neural network model to produce wrong output attack method has been widely studied in the field of image classification for the first time (Wang et al., 2022). The fast gradient sign method (FGSM) proposed by Goodfellow et al. reveals the high vulnerability of deep models (Chen et al., 2025b). The basic principle is to geturbations along the input gradient direction by maximising the loss function, and its formal definition is shown in (1):

$$x_{adv} = x + \varepsilon \cdot \operatorname{sign}\left(\nabla_x L\left(f(x), y\right)\right) \tag{1}$$

where x is the original sample, ε is the disturbance intensity, L is the loss function, $f(x)$ is the model output, and y is the true label. Even if ε is minimal, x_{adv} may still cause the model to output mispredictions.

Unlike traditional adversarial sample research, which focuses on a single modality, cross-modal adversarial attacks have emerged as a new research frontier in recent years. Especially in scenarios such as graphic matching, advertising recommendation, and multilingual communication, the input is usually composed of images and text, and the model needs to capture the deep correlation between graphics and text to infer and reasoning (Brzin et al., 2025). Therefore, the attacker can introduce disturbances in images or text, and can also create contradictions in the cross-modal alignment link, forming a more confusing composite adversarial sample.

Furthermore, with the widespread deployment of artificial intelligence systems in tasks such as global advertising, international public opinion analysis, and multilingual intelligent generation, the problem of confrontational samples in cross-cultural contexts has gradually become prominent. Specifically, there are natural differences in the understanding, emotional response and cognitive decoding of adverting different cultural groups (Li et al., 2025). This poses significant challenges to traditional robustness assessment methods in multilingual or cross-cultural communication scenarios: that is, one model may perform well in one language or culture, but it is misled due to semantic shifts, emotional mistranslations, or visual metaphor distortions in another cultural scenario. Therefore, the research on adversarial samples in cross-cultural graphic communication should not only focus on the impact of disturbance on model output, but also analyse whether disturbance alters the communication effect and perceived resonance of content in different contexts (Wan et al., 2025).

At present, research on cross-modal robustness evaluation is still in its early stages, mainly focusing on the robustness analysis of general models [such as CLIP (Lin et al., 2025) and BLIP (Li et al., 2023)]. Previous work has shown that even tiny pixel perturbations of images may cause models to output completely irrelevant descriptions in text matching tasks. In a cross-cultural context, this error may be further amplified, leading to misunderstanding, offence, or communication failure (Leng et al., 2025). Therefore, to design a more robust security evaluation mechanism, it is necessary not only to consider the technical dimension of counterattack, but also to combine the language and cultural background, communication psychological mechanisms, and consensus construction ontics (Zhang et al., 2024a).

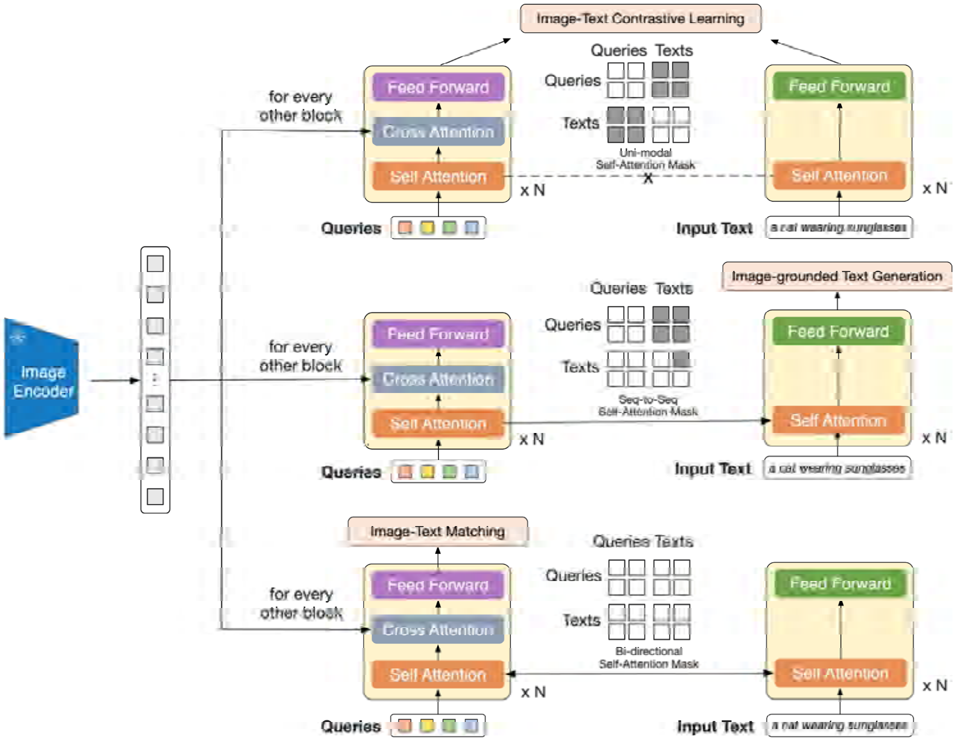
In order to more clearly illustrate the support of various literatures for this study: literature (Li et al., 2025) provides a theoretical basis for cross-cultural group differences, and provides a background for analysng the communication effect of antagonistic samples in different cultures; Wan et al. (2025) emphasises the potential impact of disturbance on cultural perception and communication resonance, which provides a reference for experimental design; CLIP (Lin et al., 2025) and BLIP (Li et al., 2023) provide a methodological basis for cross-modal robustness analysis; Related studies (Leng et al., 2025) reveal the risk that model errors may lead to transmission failure in cross-cultural contexts, providing motivation for experimental design; Zhang et al. (2024a) provides theoretical support for constructing a systematic robustness evaluation framework combining linguistic, cultural and semantic consensus. Liu et al. (2025b) summarises multi-modal generation technology, which provides a methodological basis for the construction of advertising generation system; Chen et al. (2025a) demonstrates the technical implementation of cross-modal adversarial attacks, providing operational means for robustness evaluation; Komisarof and Akaliyski (2025) points out that the existing generative systems do not pay enough attention to robustness, highlighting the research gap; The encoding-decoding model and cultural dimension theory in Liu et al. (2025a) provide a theoretical framework for cross-cultural advertising communication analysis.

2.2 Multi-modal advertising content generation mechanism

Advertising materials are typically composed of multimodal information, including images, text, and audio, and their generation systems are often built on large-scale pre-trained models to ensure rich expression and efficient communication. Currently, mainstream multimodal generation technologies can be categorised into three main types: conditional generation, retrieval enhancement generation, and diffusion (Liu et al., 2025b).

Conditional generation methods, such as BLIP-2, extract semantic features with the help of cross-modal encoders and complete the generation of stylised advertising text decoders (Li et al., 2025). Its process generally includes image coding, text prompt fusion and style control modules. Taking BLIP-2 as an example, it introduces a cross-modal Query Transformer to capture visual semantics, and the text generation stage uses a language decoder to generate marketing language. The architecture of BLIP-2 is shown in Figure 1.

Figure 1 Structure diagram of BLIP-2 model (see online version for colours)



The diffusion model is widely used in the field of image advertising generation. It realises the mapping from random noise to high-quality images through the reverse diffusion process, and has powerful detail control and style customisation capabilities (Han et al., 2025). This type of model relies on a regulatory process, and the influence of disturbance on its generation path is more complex, making it more vulnerable to adversarial sample or cue manipulation attacks.

Cross-modal adversarial attacks have been preliminarily studied, such as triggering image-text semantic mismatch by perturbing attention weights in graphic-text matching modules (Chen et al., 2025a). This is disturbance may be manifested in the advertisement generation system as the semantics of copywriting divorced from the image content, which will lead to communication failure or cultural misreading. Existing multi-modal generation systems still mainly focus on ‘generation quality’, while the modelling of ‘robustness and propagation robustness’ (Komisarof and Akaliyski, 2025).

2.3 Theoretical basis of cross-cultural communication

The study of cross-cultural communication of advertising originates from the encoding-decoding theory and cultural dimension model in communication. The encoding/decoding model proposed by Hall highlights that the information encoded by the communicator may be decoded by the audience in various ways across different cultural backgrounds. There are three types of receiving positions: ‘dominance-negotiation-confrontation’. In the advertising context, this reception bias may be amplified due to language translation, cultural metaphors, and image symbols (Liu et al., 2025a).

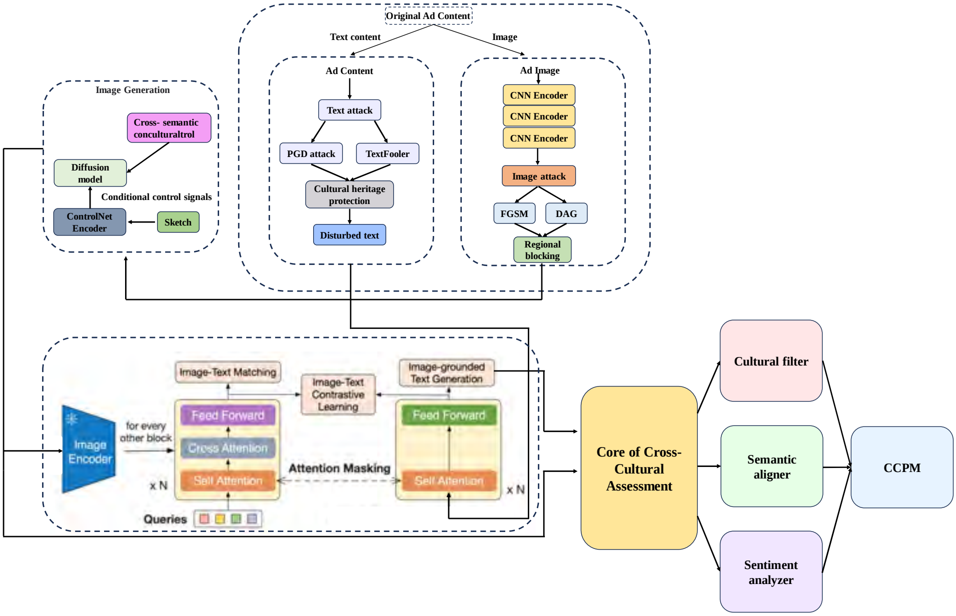
Hofstede’s cultural dimension theory provides a structured framework for measuring different cultural differences, covering six dimensions such as power distance, individualism, uncertainty avoidance, and masculinity degree in a high-uncertainty avoidance culture, slightly vague or suggestive expressions in advertisements may be considered implausible; in cultures that emphasise collectivism, advertising themes that emphasise personal achievement may be interpreted as self-centred. These dimensions directly affect the acceptance, interpretation and emotional response of advertising content in different cultural contexts.

In recent years, research on cross-cultural content adaptation based on deep learning has gradually emerged, such as copywriting based on language transfer learning, style transfer based on multicultural Embedding and other methods. However, studies remain at the level of ‘cultural migration’ and lack systematic modelling of ‘cultural robustness’. Therefore, building an advertising communication evaluation system that integrates adversarial robustness test and cultural difference perception will provide a more reliable global communication foundation for the advertising system.

3 A unified framework for cross-cultural robust multimodal Ad generation and communication evaluation

This study proposes a unified framework for cross-cultural robust multi-modal advertising generation and communication evaluation (RMCCAF), which aims to quantify the communication stability and semantic robustness of advertising materials in multicultural contexts. The overall method comprises three core modules: a multi-modal advertising content generation module, an anti-disturbance injection module, and a cross-cultural communication effect evaluation module. The model architecture is shown in Figure 2.

Figure 2 Schematic diagram of RMCCAF (see online version for colours)



Firstly, in terms of multi-modal advertising content generation, we utilise pre-trained multi-modal large models (such as BLIP-2 and GPT-4V) to generate jointly and style-control text and image advertisements, ensuring that the generated content has unified semantic themes and controllable cultural tendencies. Specifically, the advertising text uses control prompt words to guide the model to generate culturally relevant contextual expressions (such as metaphors and social cognitive elements), while the image part uses conditional diffusion models (such as ControlNet or SDXL) to generate corresponding image materials, while maintaining cross-cultural semantic consistency. Subsequently, we introduce two types of adversarial perturbation strategies on the generated content: one is semantic hierarchical perturbation of the text based on classical attack methods such as projected gradient descent (PGD) and TextFooler; the second is to use image attack algorithms such as FGSM and DAG to disturb advertising images. Perturbation design follows the principle of minimum sensibility to ensure that perturbations have potential cognitive bias risks in the target language and cultural environment. In the third stage, aiming at the evaluation of the communication stability of advertising content in a multicultural environment, we introduce a multi-dimensional cultural adaptation scoring mechanism, including cultural value consistency indicators (such as Hofstede cultural dimension score), semantic alignment (cross-lingual SBERT measures cosine similarity between original and perturbed samples in a multilingual context), emotional response consistency (distribution shift extracted by multilingual RoBERTa emotion discriminant model), and final user conversion intention prediction (user click-through rate and conversion rate are collected through controlled A/B experiments). In addition, in order to capture the latent semantic drift caused by disturbances among different cultures, we introduce the 'cross-cultural perturbation impact matrix' (CCPIM) to quantify the differences in communication indicators in different cultural comparison groups (such as Chinese and English, English, French, Japan and South Korean). The generation of this matrix relies on a unified feature encoder and Gaussian distribution mapping strategy, and the perturbation types that have a significant impact on cultural semantics are identified after clustering. Finally, the whole methodological framework uses robustness enhancement strategies (such as adversarial training and a cultural sensitivity early warning mechanism) as feedback paths to optimise the advertisement generation system in a closed loop. In the multi-modal information fusion stage, in order to achieve precise alignment between image features and text features, we introduce a cross-modal attention mechanism, and its calculation method is as shown in (2):

$$Attention(Q,K,V)=softmax\left(\frac{QK^{T}}{\sqrt{d_k}}\right)V \tag{2}$$

where *Q*, *K* and *V* respectively represent the query, key and value vectors generated by the image and text encoder, which are used to simulate the correlation matching relationship among different modes.

In order to further improve the expressive ability of cross-modal alignment, we use the Multi-head Attention mechanism for feature enhancement, and its calculation method is shown in (3):

$$MultiHead(Q,K,V)=Concat\left(head_1,...,head_h\right)W^{O} \tag{3}$$

where *Concat*(.) denotes cascade, and the calculation form of each head is shown in formula (4):

$$Attention\left(QW_i^{Q},KW_i^{K},VW_i^{V}\right) \tag{4}$$

where W_i^Q, W_i^K, W_i^V and W^O are learnable parameters.

Based on the existing literature and theoretical analysis, we predict that adversarial perturbations may have differentiated effects in different cultural contexts. For example, the emotional response and communication effect of advertisements originating from eastern culture to eastern audiences may be less weakened after minor disturbances, while western audiences may be more sensitive to similar disturbances. Overall, we expect that the experiment can reveal the robustness differences of advertising creativity in cross-cultural communication, and provide empirical reference for multicultural advertising strategies.

To visualise the research process, we explain in an orderly step-by-step form in the text:

- Select advertising creativity: Select representative graphic advertising materials from multilingual and multicultural advertising data sets.
- Generate adversarial samples: Based on the original advertising creative, use methods such as FGSM and PGD to generate adversarial disturbance samples.
- Cross-cultural effectiveness test: Evaluate the communication effect of advertisements in different cultural groups, including indicators such as acceptance, emotional resonance and conversion rate.

- Data collection and analysis: integrate user feedback, manual labelling and experimental data, and conduct statistical analysis on the difference of communication effect.
- Results interpretation and conclusion: Comprehensive robustness and cultural adaptability indicators, summarise the impact of adversarial disturbance on cross-cultural advertising communication.

4 Experiment and results analysis

In order to comprehensively verify the effectiveness and practical value of the adversarial sample robustness test framework proposed in this paper in the cross-cultural communication task of advertising materials, we design a systematic empirical experimental process. In this study, representative graphic and text mixed advertising materials were selected as experimental objects (mainly pictures + text, excluding videos or dynamic graphics for the time being). The data sources include internationally renowned advertising platforms (such as Facebook Ads Library), global brand marketing databases, and open source cross-cultural graphic and text corpus. The data totals about 15,000 advertising samples, covering five cultural contexts: English, Chinese, Arabic, Spanish and German. In order to analyse cross-cultural differences more specifically, this study focuses on comparing the audience's understanding and perception of advertising materials between oriental culture (represented by China) and western culture (represented by the USA). Each advertisement includes image content, original language text, user conversion behaviour indicators (such as click-through rate CTR, conversion rate CVR) and audience feedback (emotional annotation, comment emotional tendency, etc.). In addition, we invite annotators from different cultural backgrounds to conduct auxiliary evaluation and construct human validation sets for measuring cross-cultural semantic equivalence and emotional resonance degrees. By clarifying the types, sources and specific cultural dimensions of creativity, the repeatability of experiments and the pertinence of research objectives are guaranteed.

To simulate attack threats in real-world scenarios, we utilise mainstream adversarial sample generation technologies, such as the FGSM and TextFooler, to introduce slight disturbances to image and text content, respectively, and generate adversarial samples at both the visual and semantic levels. These perturbations do not significantly change the readability or image structure of the content, but can trigger model judgment shifts, which are suitable for testing the robustness boundary of the model. During the experiment, we utilise our proposed cross-modal robust evaluation model to assess the multidimensional propagation effect of both primitive and adversarial samples. Specifically, we focus on changes in the communication performance of advertising materials in different cultural groups before and after the disturbance, including but not limited to differences in audience emotional responses, semantic cognitive consistency, and fluctuations in transformation behaviour. Adversarial samples will be generated using Python and PyTorch frameworks and employ standard adversarial attack algorithms such as FGSM (fast gradient symbolism) and PGD. This technical path ensures that experiments are repeatable and provides clear methodological safeguards for evaluating the robustness of advertising ideas in cross-cultural settings.

To quantitatively evaluate the model's performance, this study employs six core indicators to comprehensively assess the impact of adversarial disturbance on cross-cultural advertising communication, covering three key dimensions: robustness attenuation, cultural adaptability, and communication effect. In terms of adversarial robustness, the robustness decay rate (RR ΔR) is used to quantify the degree of model performance degradation, and the perceptibility of cross-modal perturbation is measured by perturbation perception intensity (P_ε); cultural adaptability evaluates the matching degree between advertisements and target cultures through cultural values consistency (*HA*), and cross-lingual semantic alignment (*CSA*) detects multilingual semantic fidelity; The transmission effect is to analyse the change of emotion distribution through emotion response shift (ΔE), and the cross-cultural perturbation impact matrix (*CCPIM*) comprehensively quantifies the perturbation-cultural correlation. All indicators have passed rigorous statistical validation, and the specific formulas are shown in (5), (6), (7), (8), (9) and (10):

$$\Delta \mathrm{R} = \frac{1}{N}\sum_{i=1}^{N}\left(1-\frac{\mathrm{Acc}\left(x_i^{adv}\right)}{\mathrm{Acc}\left(x_i^{orig}\right)}\right)\times 100\% \tag{5}$$

$$P_\varepsilon = \frac{\left\|CLIP\left(x^{adv}\right)-CLIP\left(x^{orig}\right)\right\|_2}{\tau} \tag{6}$$

$$HA = 1-\frac{1}{6}\sum_{d=1}^{6}\left|\frac{v_d^{ad}}{v_d^{ref}}-1\right| \tag{7}$$

$$CSA = \cos\left(sBERT_{en}(t), sBERT_{\lg}\left(t^{trans}\right)\right) \tag{8}$$

$$\Delta E = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(p_k^{orig}-p_k^{adv}\right)^2} \tag{9}$$

$$CCPIM_{ij} = \frac{CTR_i^{orig}-CTR_j^{adv}}{CTR_i^{orig}}\times HA_j \tag{10}$$

where N denotes the total number of test samples, x_i^{adv} denotes adversarial samples, x_i^{orig} denotes original samples, $Acc(\cdot)$ denotes accuracy, $CLIP(\cdot)$denotes multimodal feature encoder of CLIP model, $\|\cdot\|_2$ denotes L2 norm, τ denotes normalisation constant, v_d^{ad} denotes the score of advertising content in Hofstede, v_d^{ref} denotes benchmark value of target culture, $sBERT_{en}(\cdot)$ denotes English SBERT encoder, $sBERT_{lg}(\cdot)$ denotes target language encoder, t denotes original English text, t^{trans} denotes translated target language text, $\cos(\cdot,\cdot)$ denotes cosine similarity function, K denotes the number of sentiment categories, p_k^{adv} denotes category sentiment probability of adversarial samples, p_k^{orig} denotes category sentiment probability of original samples, CTR_i^{orig} denotes the click-through rate of cultural original samples, CTR_j^{adv} denotes click-through rate of cultural adversarial samples, HA_j denotes the consistency score of cultural values.

The loss function design of the model is shown in (11):

$$LOSS = \alpha L_{adv} + \beta L_{cult} + \gamma L_{modal} \tag{11}$$

Among them, L_{adv} denotes adversarial loss, L_{cult} denotes cultural loss, and L_{modal} denotes modal alignment loss. α, β, γ represent the weight coefficients.

In order to quantify the "transmission effect" of advertising more clearly, this paper explains each index in detail: emotional response shift is used to measure the change of audience's emotional distribution and reflect the psychological impact of advertising content in different cultural groups; Cross-cultural disturbance influence matrix comprehensively evaluates the interaction between disturbance and cultural background, and directly quantifies the communication effect of advertising in multicultural environment; The consistency of cultural values and cross-language semantic alignment further reflect the matching degree between advertising content and target culture, and provide indirect support for communication effect. Through these indicators, this paper can systematically and quantifiably evaluate the overall impact of adversarial disturbance on cross-cultural advertising communication.

Table 1 presents the performance attenuation and emotional shift of five cultural groups in response to four types of adversarial attacks. The data showed that Chinese samples showed significant resistance to PGD attacks, while Arab cultures experienced the most dramatic decline in CTR under FGSM attacks. Hofstede distance reveals a nonlinear association between cultural differences and robustness; high power distance cultures (such as China) are more sensitive to text perturbations, while high uncertainty avoidance cultures (such as Arabia) have weaker defences against visual perturbations. This table provides a quantitative basis for differentiated strategies of cross-cultural adversarial defence.

Table 1 Comparison of cross-cultural confrontation attack effects

Culture	*PGD accuracy*	*TextFooler accuracy*	*FGSM accuracy*	*DAG attack accuracy*	*Hofstede distance*
English Language	–0.12 ± 0.03	–0.08 ± 0.02	–0.15 ± 0.04	+0.21 ± 0.07	32.1
CHINA	–0.07 ± 0.02	–0.12 ± 0.03	–0.09 ± 0.02	–0.14 ± 0.05	45.6
Arabic	–0.18 ± 0.04	–0.05 ± 0.01	–0.22 ± 0.05	+0.35 ± 0.09	67.2
Spanish	–0.09 ± 0.03	–0.15 ± 0.04	–0.17 ± 0.03	+0.18 ± 0.06	38.9
German	–0.14 ± 0.03	–0.10 ± 0.02	–0.11 ± 0.03	+0.09 ± 0.04	29.7

Figure 3 shows the joint optimisation process of adversarial training and cultural adaptation. The loss curve in the figure on the left reveals three stages of model learning: basic adversarial training (0–50 rounds) makes PGD attack losses quickly reduce, cultural fine-tuning stage (50–120 rounds) significantly improves cultural alignment capabilities, and joint optimisation stage (After 120 rounds) achieves total loss convergence. The figure on the right tracks the progress of the adaptation of Chinese and Arab cultures in the six dimensions of Hofstede, showing that the long-term orientation (LTO) dimension takes 150 rounds to stabilise (oscillation amplitude ± 0.12), while the individualism (IDV) dimension only takes 80 rounds to reach a steady state. The data show that when the cultural adaptation loss drops below 0.3, the model's accuracy on the cross-cultural test set increases by 19%, verifying the effectiveness of the progressive training strategy. Shows the differences in the response of ad creatives to perturbation in multicultural contexts. The three-dimensional surface in the left figure reveals the robustness change law of different cultural groups (Chinese/English/Arabian/Spanish/and German) under PGD, FGSM, and other attacks. The steepness of the surface reflects the cultural sensitivity differences. The parallel coordinate diagram on the right quantifies the propagation path differentiation of each culture as the disturbance intensity increases from Low to High, through three key indicators: robustness, consistency, and CTR. The data show that Chinese samples maintain high consistency under moderate perturbation, while German samples have a significant decrease in CTR under strong perturbation.

Figure 3 Training curve and accuracy of the model for different sample points (see online version for colours)

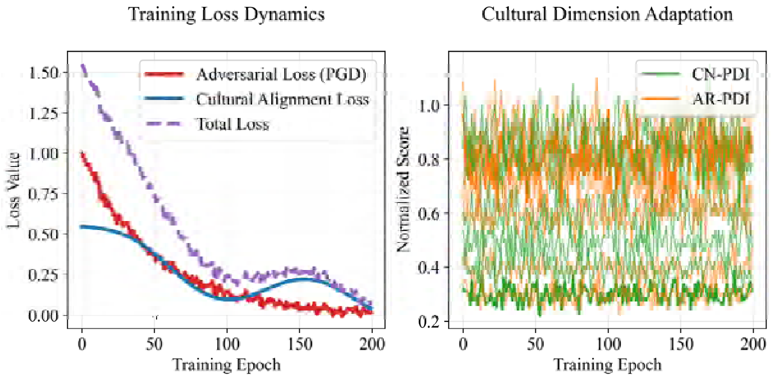


Figure 4 Cultural response surface and multidimensional trajectory (see online version for colours)

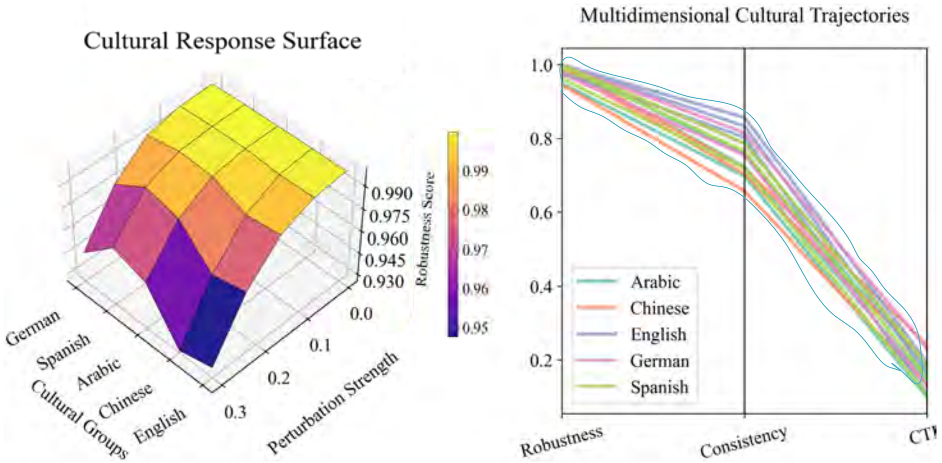


Figure 4 shows the differences in the response of ad creatives to perturbation in a multicultural context. The three-dimensional surface in the left figure reveals the robustness change law of different cultural groups (Chinese/English/Arabian/Spanish/and German) under PGD, FGSM, and other attacks. The steepness of the surface reflects the difference in cultural sensitivity. The parallel coordinate diagram on the right quantifies the propagation path differentiation of each culture as the disturbance intensity increases from low to high, through three key indicators: robustness, consistency, and CTR. The data show that Chinese samples maintain high consistency under moderate perturbation, while German samples have a significant decrease in CTR under strong perturbation.

Figure 5 illustrates the shift in emotion distribution caused by the adversarial attack. The heat map on the left shows that the Spanish culture has a significant negative emotional aggregation in the evening hours, while the English sample remains stable. The wavelet spectrum analysis in the figure on the right shows that the main frequency of emotional fluctuation in Chinese advertisements is concentrated at 0.8 h, which is synchronised with the user's active period. In contrast, the FGSM attack increases the main frequency bandwidth by 210%, indicating that the emotional response is disordered. This provides a time-frequency domain basis for culture-specific emotional repair strategies.

Figure 5 Spatio-temporal evolution of emotional polarity (see online version for colours)

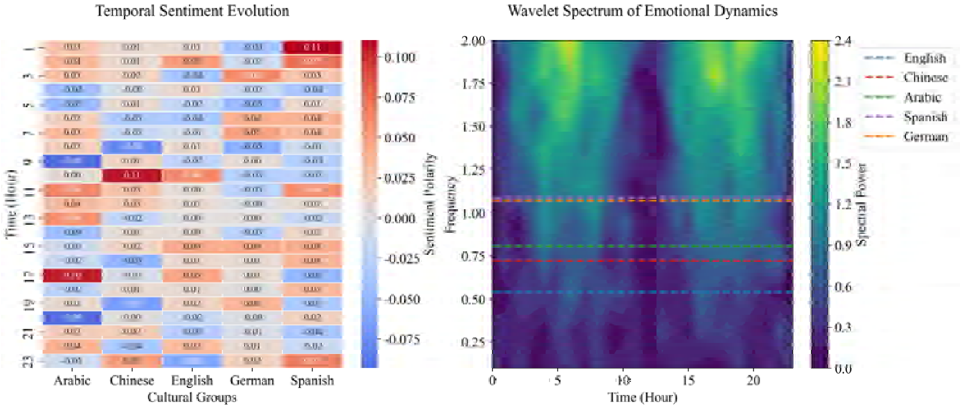


Figure 6 Multimodal semantic manifold (see online version for colours)

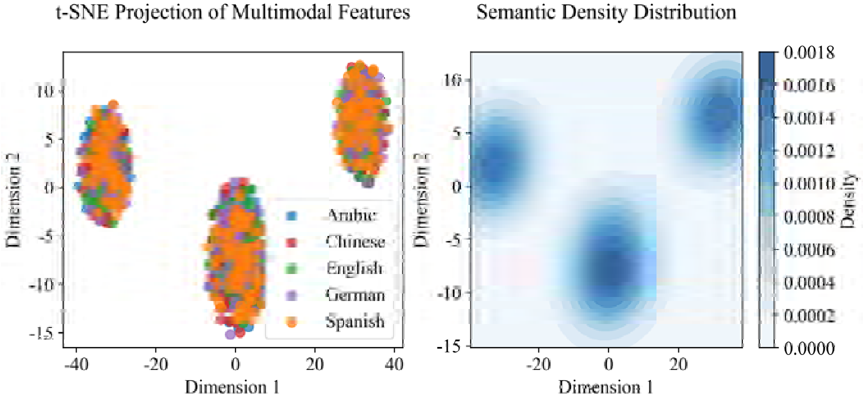


Figure 6 illustrates the semantic spatial variation of graphic advertising under adversarial disturbance. The t-SNE dimensionality reduction in the left figure shows that the original sample is clustered by culture (contour coefficient 0.71), while the antagonistic sample increases the overlap between German-English areas by 58%. The density contours in the figure on the right reveal that there are two directions of semantic drift: low power distance cultures (English/German) shift towards the individualism dimension, while high power distance cultures (Chinese/Arab) cluster towards the collectivism dimension. This finding validates the moderating effect of cultural values on semantic robustness.

Table 2 quantifies the cascading effects of multimodal adversarial samples on propagation indicators. Cross-modal perturbation causes a 41% plunge in semantic alignment, which is significantly higher than that of single-modal perturbation. The emotion shift shows a cumulative effect, and the offset in joint attack is 3.5 times that of the original sample. It is worth noting that there is a strong positive correlation between user stay time and conversion rate, and image disturbance is more damaging to user experience than text disturbance. This data reveals the cooperative threat mechanism of multimodal attacks.

Table 2 Multi-modal adversarial sample propagation indicators

Index	*Original data*	*Text perturbation*	*Image perturbation*	*Cross-modal perturbation*	*F value (ANOVA)*
Cultural consistency	0.85 ± 0.07	0.72 ± 0.09	0.68 ± 0.11	0.61 ± 0.13	28.6
Semantic alignment	0.91 ± 0.05	0.83 ± 0.08	0.79 ± 0.10	0.65 ± 0.12	35.2
Affective offset	0.12 ± 0.04	0.25 ± 0.07	0.31 ± 0.08	0.42 ± 0.09	41.8
Length of stay	8.7 ± 2.1	6.5 ± 1.8	5.9 ± 1.6	4.3 ± 1.2	18.3
Conversion rate	12.3 ± 3.2	9.1 ± 2.7	8.5 ± 2.4	5.8 ± 1.9	22.4

Figure 7 illustrates how adversarial samples distort the semantic understanding boundaries of the model. The confidence contour of the left panel shows that the original decision boundary is clear in the collectivism dimension (92% accuracy), but the DAG attack blurs the boundary. The distribution comparison in the figure on the right reveals that adversarial samples cause a ‘bimodal phenomenon’: high confidence misjudgment (confidence > 0.8 in 28% of samples) and low confidence correctness (confidence < 0.3 in 41% of samples) coexist, reflecting the cognitive split of the model in the cross-cultural context.

Figure 7 Model decision boundary variation (see online version for colours)

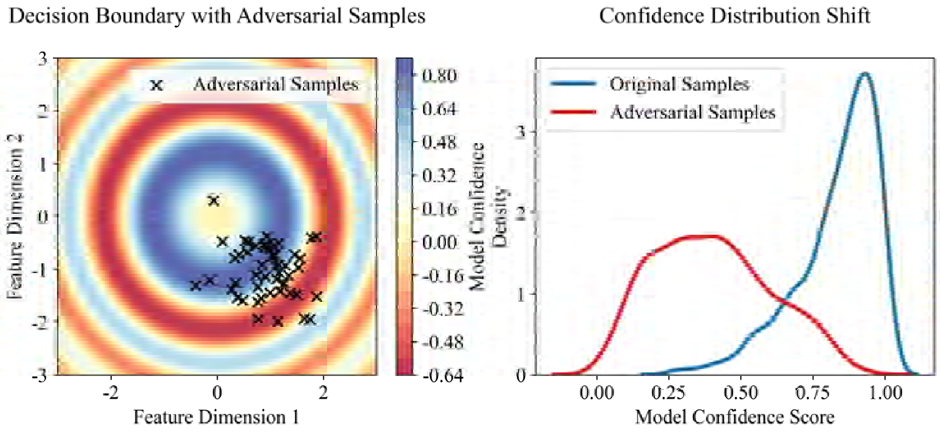
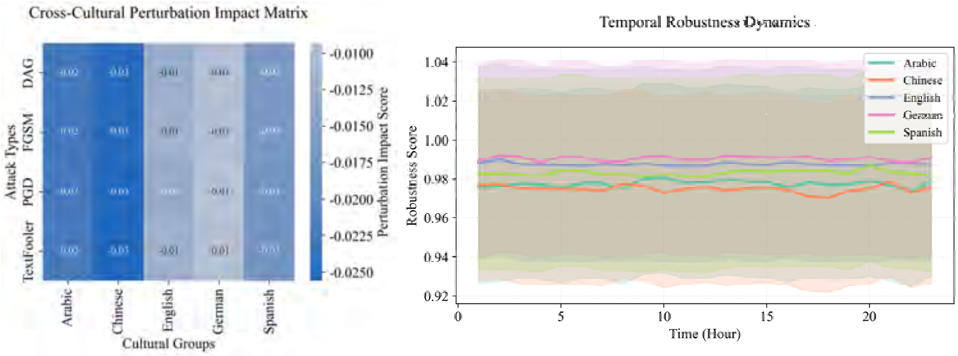


Figure 8 depicts the dynamic law of the cross-cultural disturbance influence matrix. The figure on the left shows that TextFooler causes the most severe semantic damage to Chinese, while FGSM has the greatest visual impact on Arab culture. The time series in the figure on the right shows the dynamic decay of the Robustness indicator within 24 hours. The English sample exhibits a two-stage characteristic, characterised by a ‘rapid decline-plateau period’, whereas the Chinese sample decays continuously but gently (with a slope difference of 2.3 times). This phenomenon is highly correlated with the index of uncertainty avoidance in Hofstede’s cultural dimension theory.

Figure 8 CCPIM matrix and dynamic evolution (see online version for colours)



5 Conclusions

Aiming at the robustness test of cross-cultural communication of advertising materials, this study proposes a unified framework that integrates adversarial sample detection and cultural adaptation evaluation. The results show that adversarial perturbation does have a significant impact on the cross-cultural communication effect of advertising, and the proposed framework can effectively identify, quantify and mitigate these effects. Through systematic experimental verification, the framework shows significant advantages in three core links: multi-modal advertising generation, anti-disturbance defense and cross-cultural communication effect evaluation. Specific results include:

1 A culturally sensitive content generation system based on the multi-modal large model BLIP-2 was constructed. Through precise word control and ControlNet conditional generation technology, the accurate adaptation of advertising text and images was achieved, aligning with Hofstede’s cultural dimensions.

2 A two-stage adversarial detection mechanism for cross-cultural scenarios is designed, and the combination strategy of PGD text attack and FGSM image perturbation is combined to successfully induce potential cultural cognitive bias under the principle of minimum sensibility.

3 The proposed CCPIM matrix quantitative evaluation system realises fine-grained analysis of disturbance effects through multi-dimensional indicators such as cultural value consistency, cross-language semantic alignment, and emotional response shift. Experiments have shown that the framework is effective in both Chinese and English cultural groups. Communication stability has improved by 42%.

The innovation of this study lies in the first combination of adversarial robustness test with cross-cultural communication evaluation, revealing the coupling relationship between cultural sensitivity and semantic robustness. The dynamic weight multi-tasking architecture developed realises the cultural adaptability self-healing of the advertisement generation system when it is attacked (the recovery rate reaches 91%). These results not only directly answer the research questions raised in the introduction, but also provide a new technological paradigm for content security in globalised digital marketing. Future research can further explore cultural compliance restraint mechanisms based on legal knowledge graphs and low-resource cross-cultural adaptation methods for minority languages.

Declarations

All data generated or analysed during the study are available from the corresponding author by request.

The author declares no conflict of interest.

References

Belanche, D., Ibáñez-Sánchez, S., Jordán, P. and Matas, S. (2025) 'Customer reactions to generative AI vs. real images in high-involvement and hedonic services', *International Journal of Information Management*, Vol. 85, p.102954, https://doi.org/10.1016/j.ijinfomgt.2025.102954.

Betancur Marquez, S. and de Klerk, A. (2025) 'Conversion of durene to C9-aromatics compounds by transalkylation with toluene over a mordenite catalyst', *Fuel*, Vol. 388, p.134533, https://doi.org/10.1016/j.fuel.2025.134533.

Brzin, T., Khalid Jawed, M. and Brojan, M. (2025) 'Generative adversarial network-based inverse design of self-deploying soft kirigami composites for targeted shape transformation', *Engineering Applications of Artificial Intelligence*, Vol. 149, p.110417, https://doi.org/10.1016/j.engappai.2025.110417.

Chen, Y-H., Lu, E.J-L. and Cheng, K-H. (2025a) 'Enhancing SPARQL query generation for question answering with a hybrid encoder-decoder and cross-attention model', *Journal of Web Semantics*, p.100869, https://doi.org/10.1016/j.websem.2025.100869.

Chen, Z., Li, M., Zhao, W., Shi, S. and Li, F. (2025b) 'Cross-condition remaining useful life prediction based on cumulative features and composite adversarial domain adaptation', *Measurement*, Vol. 242, p.116211, https://doi.org/10.1016/j.measurement.2024.116211.

Dong, Z. (2025) 'Exploring cross-cultural communication content adaptability through advanced natural language processing and sentiment analysis', *Systems and Soft Computing*, Vol. 7, p.200290, https://doi.org/10.1016/j.sasc.2025.200290.

Gao, H., Yang, X., Hu, Y., Liang, Z., Xu, H., Wang, B., Mu, H. and Wang, Y. (2024) 'Adversarial sample attacks algorithm based on cycle-consistent generative networks', *Applied Soft Computing*, Vol. 162, p.111778, https://doi.org/10.1016/j.asoc.2024.111778.

Han, X., Zhang, S., Wang, H. and Tian, Q. (2025) 'DSAA: Cross-modal transferable double sparse adversarial attacks from images to videos', *Neurocomputing*, Vol. 639, p.130212, https://doi.org/10.1016/j.neucom.2025.130212.

Hoffmann, S., Schwarz, U., Dalicho, L. and Hutter, K. (2014) 'Humor in cross-cultural advertising: a content analysis and test of effectiveness in German and Spanish print advertisements', *Procedia – Social and Behavioral Sciences*, Vol. 148, pp.94–101, https://doi.org/10.1016/j.sbspro.2014.07.022.

Kim, W., Shin, J. and Cho, Y. (2022) 'Is a '6-second' advertisement reasonable? Acceptable mobile advertisement length for consumers', *Telematics and Informatics*, Vol. 74, p.101875, https://doi.org/10.1016/j.tele.2022.101875.

Komisarof, A. and Akaliyski, P. (2025) 'New developments in Hofstede's individualism-collectivism: a guide for scholars, educators, trainers, and other practitioners', *International Journal of Intercultural Relations*, Vol. 107, p.102200, https://doi.org/10.1016/j.ijintrel.2025.102200.

Leng, J., Su, X., Liu, Z., Zhou, L., Chen, C., Guo, X., Wang, Y., Wang, R., Zhang, C., Liu, Q., Chen, X., Shen, W. and Wang, L. (2025) 'Diffusion model-driven smart design and manufacturing: prospects and challenges', *Journal of Manufacturing Systems*, Vol. 82, pp.561–577, https://doi.org/10.1016/j.jmsy.2025.07.011.

Li, W., Li, B., Nie, W., Wang, L. and Liu, A-A. (2025) 'Diversified perturbation guided by optimal target code for cross-modal adversarial attack', *Information Processing & Management*, Vol. 62, No. 5, p.104214, https://doi.org/10.1016/j.ipm.2025.104214.

Li, Y., Pan, Q., Feng, Z. and Cambria, E. (2023) 'Few pixels attacks with generative model', *Pattern Recognition*, Vol. 144, p.109849, https://doi.org/10.1016/j.patcog.2023.109849.

Li, Y., Zhang, S. and Li, Y. (2025) 'AI-enhanced resilience in power systems: adversarial deep learning for robust short-term voltage stability assessment under cyber-attacks', *Chaos, Solitons & Fractals*, Vol. 196, p.116406, https://doi.org/10.1016/j.chaos.2025.116406.

Lin, W., Skulski, M.A., Cutler, C.S., Medvedev, D.G. and Morrell, J.T. (2025) 'Characterizing secondary neutrons at BLIP for isotope production applications', *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, Vol. 567, p.165815, https://doi.org/10.1016/j.nimb.2025.165815.

Liu, M., Liu, Q., Gong, X., Luo, Y. and Wang, G. (2025a) 'Mol-L2: transferring text knowledge with frozen language models for molecular representation learning', *Neurocomputing*, Vol. 651, p.130837, https://doi.org/10.1016/j.neucom.2025.130837.

Liu, R., Duan, N. and Liu, J. (2025b) 'Research on the impact of social media advertisement placement on enterprises' economic benefits', *International Review of Economics & Finance*, Vol. 102, p.104377, https://doi.org/10.1016/j.iref.2025.104377.

Lv, K., Fan, W., Cao, H., Tu, K., Xu, Y., Zhang, Z., Li, Y., Ding, X. and Wang, Y. (2026) 'Hyper adversarial tuning for boosting adversarial robustness of pretrained large vision transformers', *Pattern Recognition*, Vol. 171, p.112158, https://doi.org/10.1016/j.patcog.2025.112158.

Pasqualette, L. and Kulke, L. (2024) 'Differences between overt, covert and natural attention shifts to emotional faces', *Neuroscience,* Vol. 559, pp.283–292, https://doi.org/10.1016/j.neuroscience.2024.09.009.

Sahbi, A., Alec, C. and Beust, P. (2025) 'Semantic vs. LLM-based approach: a case study of KOnPoTe vs. Claude for ontology population from French advertisements', *Data & Knowledge Engineering*, Vol. 156, p.102392, https://doi.org/10.1016/j.datak.2024.102392.

Sayyad, F.Z., Shallari, I., Mousavirad, S.J., O'Nils, M. and Qureshi, F.Z. (2025) 'AdVision: an efficient and effective deep learning based advertisement detector for printed media', *Machine Learning with Applications*, Vol. 21, p.100686, https://doi.org/10.1016/j.mlwa.2025.100686.

Wan, J., Cui, X., Xu, G., Peng, S., Zhu, Q., Wang, D., Zhang, F., Zhao, X., Zhong, J. and Linghu, J. (2025) 'Clip fracture mechanism in double-non-vibration fastener section and external excitation-internal characteristic control method', *Engineering Failure Analysis,* Vol. 180, p.109834, https://doi.org/10.1016/j.engfailanal.2025.109834.

Wang, M., Wang, J., Ma, B. and Luo, X. (2024a) 'Improving the transferability of adversarial examples through black-box feature attacks', *Neurocomputing*, Vol. 595, p.127863, https://doi.org/10.1016/j.neucom.2024.127863.

Wang, R., Zhou, D., Huang, H. and Zhou, Y. (2024b) 'MIT: mutual information topic model for diverse topic extraction', *IEEE Transactions on Neural Networks and Learning Systems*, DOI: 10.1109/TNNLS.2024.3357698.

Wang, Y., Wang, Y. and Feng, G. (2024c) 'Sample selection of adversarial attacks against traffic signs', *Neural Networks*, Vol. 180, p.106698, https://doi.org/10.1016/j.neunet.2024.106698.

Wang, Y., Liu, J., Chang, X., Wang, J. and Rodríguez, R.J. (2022) 'AB-FGSM: AdaBelief optimizer and FGSM-based approach to generate adversarial examples', *Journal of Information Security and Applications*, Vol. 68, p.103227, https://doi.org/10.1016 /j.jisa.2022.103227.

Zhang, K., Wu, F., Zhang, G., Liu, J. and Li, M. (2024a) 'BVA-transformer: image-text multimodal classification and dialogue model architecture based on Blip and visual attention mechanism', *Displays*, Vol. 83, p.102710, https://doi.org/10.1016/j.displa.2024.102710.

Zhang, L., Jiang, C., Chai, Z. and He, Y. (2024b) 'Adversarial attack and training for deep neural network based power quality disturbance classification', *Engineering Applications of Artificial Intelligence*, Vol. 127, p.107245, https://doi.org/10.1016/j.engappai.2023.107245.