IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 42, No. 6, December, 2025, pp. 3103-3114
Journal homepage: http://iieta.org/journals/ts

Quantitative Evaluation Model of Advertising Visual Effects Based on Image Saliency and Social Attention Mechanism

Feiyang Zhang

School of Communication, Northwestern University, Evanston 60201, USA

Corresponding Author Email: zffy1021@163.com

https://doi.org/10.18280/ts.420606

Received: 22 April 2025
Revised: 17 August 2025
Accepted: 1 October 2025
Available online: 31 December 2025

Keywords:
advertising visual effect quantification, image saliency detection, social attention mechanism, multi-scale feature fusion, transformer, multi-task learning, AIGC advertising evaluation

ABSTRACT

The global digital advertising market has surpassed $600 billion, and the rapid proliferation of AIGC-generated advertisements has raised an urgent need for automated and scalable visual effect quantification tools. Traditional evaluation methods rely on manual scoring or single behavioral metrics, which suffer from strong subjectivity and limited dimensionality, while failing to integrate the two core driving factors of advertising effectiveness: visual saliency and social attention. Existing models generally lack a systematic theoretical framework, with key parameters and weights lacking scientific foundation. Moreover, global feature modeling and multi-task collaborative mechanisms are unclear, resulting in insufficient correlation between the quantification results and real user behaviors. To address these issues, this paper proposes the Saliency-Social Attention Transformer (SSAT) model: a dual-input feature system that combines visual saliency and social attention, with feature construction parameters optimized through meta-analysis and data statistics. We design a social attention-enhanced Transformer encoder to achieve collaborative modeling of local features and global context. A multi-task learning architecture is employed to simultaneously predict core advertising visual effect metrics and classify attributes. Experimental results show that the proposed model significantly outperforms general visual models and existing advertising-specific models in both visual effect quantification and attribute classification tasks, achieving a Pearson Correlation Coefficient (PCC) of 0.86 in attention capture rate and a classification accuracy of 89.2%. The core of this performance lies in the collaborative mechanism of the dual attention model, with ablation experiments demonstrating a notable decline in performance when either module is removed. The complementary nature of these two mechanisms supports the dual-channel feature capture of "visual attraction-social value." Furthermore, the model's objective quantification metrics are highly aligned with human subjective perception and exhibit good adaptability across different advertising dissemination scenarios, providing a reliable solution for automated advertising visual effect evaluation.

1. INTRODUCTION

The global digital advertising market has surpassed $600 billion, and AIGC technology [1-3] has propelled the advertising generation process into a scaled phase, with a daily output exceeding million. This industrial transformation raises new requirements for advertising effect evaluation, as traditional manual evaluation and single-metric evaluation models can no longer support the real-time generation, immediate evaluation, and rapid optimization of the industrial loop. The development of automated, high-precision quantification tools has become a critical need in the industry [4, 5]. In traditional evaluation methods, manual scoring relies on expert experience [6, 7], with a consistency coefficient lower than 0.65, and the process is costly and inefficient. Single behavioral metrics, such as click-through rates and dwell time, can only reflect superficial user feedback and cannot capture the deep causal links from visual attraction and cognitive processing to behavioral intention, failing to meet the needs of detailed evaluation. With the development of machine learning technology [8-10], some studies have attempted to use hierarchical Bayesian models to predict the attractiveness of advertisements [11]. For example, a visual scoring model based on VGG16 has achieved preliminary automation in evaluation but focuses only on single visual features without considering the impact of user attention mechanisms. Other studies have introduced general saliency detection algorithms to identify focal areas in advertisements [12, 13], which has somewhat improved evaluation accuracy but fails to adapt to the characteristics of key elements in advertising scenes, such as text and logos, and lacks scientific support for parameter design. Research related to social recognition has independently verified the positive impact of elements such as celebrity endorsements and authoritative certifications on advertising effectiveness [14, 15]. However, it has not integrated these social attention elements with visual features or achieved end-to-end quantification modeling. From a theoretical perspective, social recognition theory

points out that individual behavior is driven by group identity, and visual attention theory confirms that saliency features determine initial attention allocation. These two mechanisms together form the core driving force of advertising effects. However, existing studies have not systematically integrated these two mechanisms, nor have they provided a unified theoretical framework, resulting in limited interpretability and generalizability of quantification models. CNN-based methods excel at local feature extraction but cannot model global context dependencies [16, 17]. Although general Transformer models can capture long-range associations [18, 19], they have not been customized for social semantic elements in advertisements. Multi-task learning in quality evaluation has also only stayed at the level of simple task stacking [20, 21] without clarifying the collaborative mechanism and weight distribution logic.

In response to these industry needs and research gaps, this paper focuses on three core research questions: How to construct a systematic theoretical framework of dual attention-cognitive processing-behavioral intention, clarifying the intrinsic mechanisms by which visual saliency and social attention affect advertising effectiveness; How to scientifically design the key parameters and weight distribution of a dual-input feature system to enhance the model's interpretability and design rationality; How to optimize the Transformer encoder and multi-task collaborative mechanisms to achieve efficient modeling of multi-scale features and global context, strengthening the correlation between quantification metrics and real user behavior.

The main innovations of this paper are reflected in three aspects: In theory, we first propose a three-stage theoretical framework of dual attention-cognitive processing-behavioral intention, systematically revealing the internal logic by which visual saliency and social attention affect advertising effectiveness through perceptual, evaluative, and other cognitive processes. This framework expands the theoretical boundaries of intelligent advertising evaluation, and constructs a multi-dimensional quantification indicator system for AIGC advertising scenarios, breaking through the limitations of traditional single metrics and achieving precise characterization of the full-link effects of attraction, memory, and conversion. In terms of technology, we propose a scientifically interpretable dual-input feature construction method. We optimize the saliency map generation parameters based on meta-analysis of eye movement studies and sensitivity experiments with verification sets, and determine the weight of social attention elements by combining psychological experiments, statistics, and advertising research, significantly improving the rigor of model design. We design a social attention-enhanced Transformer encoder, strengthening the feature contribution of social semantic elements through spatial weight mapping and multi-head attention fusion mechanisms, breaking through the limitations of traditional CNN local modeling and the uniform modeling of general Transformers. We clarify the multi-task collaborative mechanism, verify the facilitative effect of auxiliary classification tasks on the main regression task through feature sharing visualization and weight sensitivity experiments, and optimize the loss function weight distribution scheme. In terms of data and applications, we construct the first AdVEE-10K dataset containing over 10,000 samples, covering three industries, five types of social attention elements, and three advertising types, providing triple annotation from eye movement experiments, questionnaires, and real-world placements. This effectively addresses the data scarcity problem in advertising effect quantification. The model can be directly applied to AIGC advertising generation platforms and intelligent advertising placement systems, achieving an end-to-end loop of generation-evaluation-optimization to meet the industry's cutting-edge application needs.

Based on social recognition theory, visual attention theory, and cognitive processing models, this paper constructs a three-stage theoretical framework for the formation of advertising visual effects: The first stage is the attention capture phase, where visual saliency features in advertisements such as text, logos, and color contrast attract the user's initial attention, forming the attention capture rate. The second stage is the cognitive processing phase, where social attention elements such as celebrity endorsements and authoritative certifications trigger positive user evaluation through the social recognition mechanism, working in synergy with saliency features to form memory point intensity. The third stage is the behavioral intention phase, where, based on the cognitive processing results, users form conversion intention scores, ultimately influencing actual conversion behavior. This framework clearly explains the complete chain of how the dual attention mechanism impacts advertising effects through cognitive processing mediation, providing a solid theoretical foundation for model design.

The subsequent sections of this paper are arranged as follows: Section 2 details the overall architecture and module design methods of the model; Section 3 introduces the experimental design, dataset construction, and evaluation metrics; Section 4 analyzes experimental results and performs visual validation; Section 5 discusses the core research findings, limitations of the study, and future directions; Section 6 summarizes the key contributions of the paper.

2. METHODS

2.1 Problem definition

The core task of this paper is to achieve multi-dimensional quantification evaluation and attribute classification of advertising visual effects based on advertising images. Given an advertisement image $I_{ad} \in R^{H \times W \times 3}$ with size $H \times W$, the model f is constructed to output two core results: first, a set of quantification metrics $Y = [AR, MI, CI]$ that depict the full-link advertising effects. Here, AR represents the attention capture rate, reflecting the advertisement's ability to attract initial user attention; MI represents the memory point intensity, indicating the user retention level of core advertisement information; and CI represents the conversion intention score, quantifying the user's tendency to take subsequent actions. Second, the advertisement attribute classification results C = [advertisement type, visual style, social attention type], which provide auxiliary feature support for the quantification task and achieve multi-task collaborative optimization. The overall architecture of the model is shown in Figure 1.

2.2 Input feature construction

The core drivers of advertising visual effects come from the collaborative effect of visual saliency and social attention. Therefore, this paper constructs a dual-input feature system, with the parameters and weights of both feature types

determined based on empirical research and data statistics, ensuring scientific validity and rationality. The process of constructing the advertisement saliency map is as follows: First, the YOLOv8 detector is used to identify the core elements in the advertisement and generate the element mask M_{elem}, precisely locating the areas that play a key role in attention capture. The original advertisement image is then input into a ResNet+U-Net model fine-tuned on the AdSaliency dataset to generate the initial saliency map S_{init}. To strengthen the saliency contribution of the core elements, a weighted formula is applied to optimize the feature distribution:

$$S_{ad}=S_{init}\odot M_{elem}+(1-M_{elem})\times S_{init}\times 0.3 \quad (1)$$

where, 0.3 is the optimal weight for non-core regions determined by the sensitivity experiment on the validation set. Finally, Gaussian smoothing with a 13×13 window is applied to reduce noise interference, and the result is normalized to the [0,1] range, yielding the final advertisement saliency map $I_{sal}\in R^{H\times W\times 3}$.

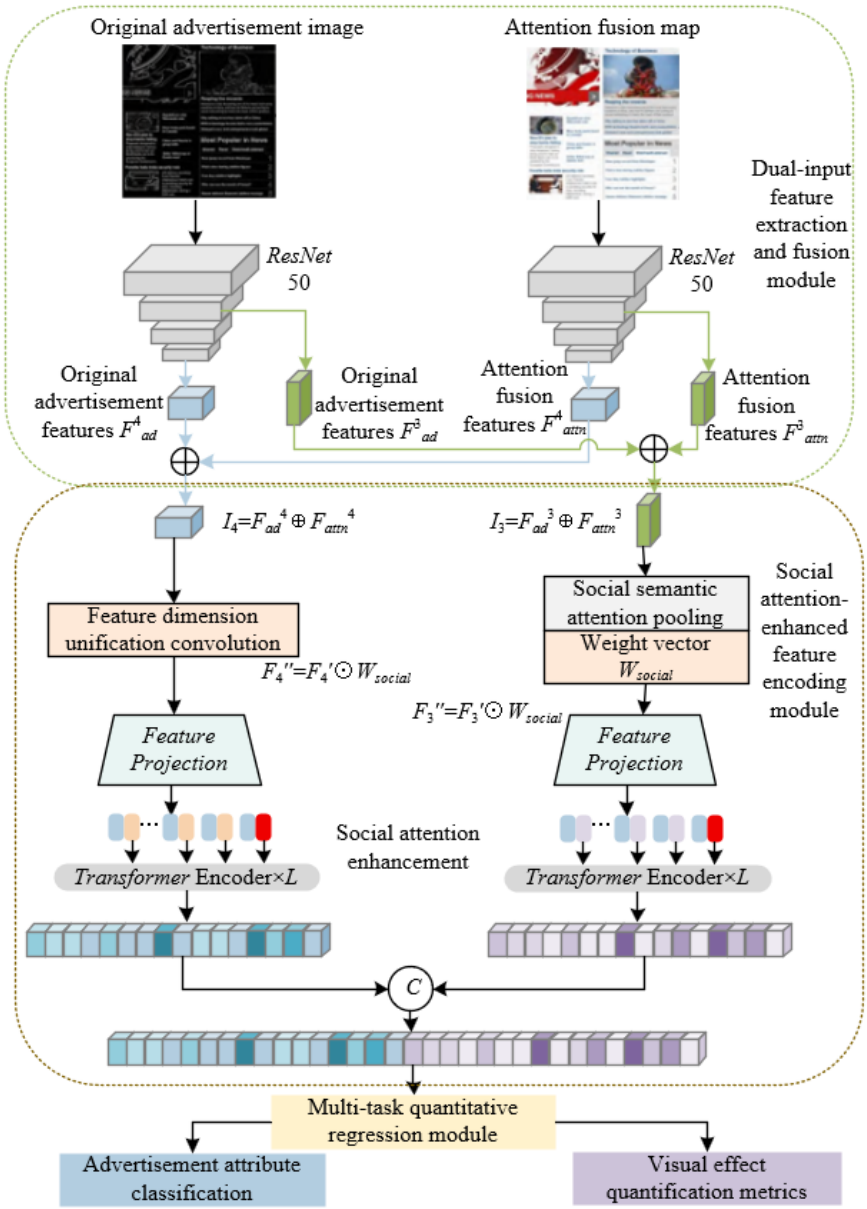


Figure 1. Overall architecture diagram of SSAT model

The construction of the social attention feature map focuses on the quantification representation of social identity elements: First, the YOLOv8 detector is used to identify the five predefined types of social attention elements. The element weights W=[0.8,0.7,0.65,0.5,0.5] are determined based on a meta-analysis of five top journal studies in psychology and advertising. Among them, the effect of celebrity endorsements has the highest weight, followed by certification marks and star ratings, while user avatars and word-of-mouth copywriting effects are similar. This weight distribution is statistically verified with annotations from the AdVEE-10K dataset. Combined with human visual center preferences, the following spatial location weight is introduced to reinforce the contribution of social elements in the image center:

$$W_{pos}=1-\frac{\text{Distance from the center in pixels}}{\max(H,W/2)} \quad (2)$$

By pixel-level multiplication, the element weight and spatial weight are fused as:

$$I_{social}(x,y)=W_{elem}\times W_{pos}(x,y) \quad (3)$$

After normalization, the social attention feature map $I_{social}\in R^{H\times W\times 3}$ is obtained, accurately characterizing the social recognition value.

2.3 Multi-scale feature extraction and fusion

To fully capture both the detailed visual information and high-level semantic features of advertisements, and to achieve deep collaboration between visual features and attention features, this paper employs a dual-branch ResNet50 architecture for multi-scale feature extraction and fusion. The first branch takes the original advertisement image I_{ad} as input, and through forward propagation of ResNet50, outputs the third and fourth-level feature maps. The third-level feature $F_{ad}^3\in R^{(H/16)\times(W/16)\times 1024}$ focuses on image texture, edges, and other detailed information, while the fourth-level feature $F_{ad}^4\in R^{(H/32)\times(W/32)\times 2048}$ represents the high-level semantics and overall structure of the advertisement. The second branch fuses the advertisement saliency map I_{sal} and social attention feature map I_{social} by pixel-wise addition to create an attention joint map I_{attn}, which is then input into another ResNet50 to obtain the corresponding attention-related features F_{attn}^3 and F_{attn}^4. To enhance the collaboration between visual features and attention features, feature fusion is achieved by pixel-wise addition across branches, i.e., $I_3=F_{ad}^3\oplus F_{attn}^3$ and $I_4=F_{ad}^4\oplus F_{attn}^4$, resulting in the fused multi-scale feature maps. These fused features provide detailed identification and semantic relevance support for subsequent global modeling.

2.4 Social attention-enhanced transformer encoder

To address the problem that traditional attention mechanisms focus mainly on physical locations or feature dimensions, and lack targeted modeling of social semantic elements, this paper designs a social attention-enhanced Transformer encoder to achieve collaborative modeling of local social semantic features and global context. In the feature preprocessing phase, the fused feature I_3 is first average pooled and downsampled to the resolution of I_4 to ensure spatial consistency across different scale features. Then, a 1×1 convolution is applied to unify the channel dimension of the two feature types and project them to D=1024, followed by flattening into 2D feature matrices $F_3'\in R^{N\times D}$ and $F_4'\in R^{N\times D}$, where $N=(H/32)\times(W/32)$ is the feature sequence length. In the social attention weight mapping step, the social attention feature map I_{social} is downsampled to the corresponding resolution and flattened into a weight vector $W_{social}\in R^{N\times 1}$, which is normalized to the [0,1] range through a Sigmoid activation function to achieve accurate mapping from social semantic elements to feature weights. Feature enhancement is then performed by element-wise multiplication: $F_3''=F_3'\odot W_{social}$ and $F_4''=F_4'\odot W_{social}$, this enhances the contribution of features with social recognition value. The enhanced feature sequences are then input into a multi-head self-attention module to model long-distance dependencies. Additionally, learnable global feature encoding and position encoding are introduced. After 6 layers of Transformer encoder iteration, the two feature types are concatenated into a global feature vector $G=[G_3,G_4]\in R^{1\times 2D}$, providing comprehensive feature representation for subsequent quantification evaluation and classification tasks.

2.5 Multi-task quantification regression module

The multi-task quantification regression module takes the global feature vector G output by the Transformer encoder as input and simultaneously performs core advertising visual effect metric prediction and attribute classification, achieving collaborative optimization through feature sharing between tasks. The module architecture is shown in Figure 2. The core task focuses on accurate regression of three quantification metrics, and a three-layer fully connected network is used to construct the prediction network. The input dimension is 2D = 2048, which is mapped to a 1024-dimensional hidden layer and outputs a 3-dimensional prediction result:

$$\hat{Y}_{metric}=[\widehat{AR},\widehat{MI},\widehat{CI}] \quad (4)$$

For the different value ranges of the metrics, differentiated activation functions are used: the attention capture rate (AR) ranges from [0,1], and a Sigmoid activation function is used to generate probabilistic outputs; the memory point intensity (MI) and conversion intention score (CI) range from [0,5], and a ReLU activation function is used to avoid gradient vanishing, ensuring the predicted values align with the actual quantification ranges of the metrics.

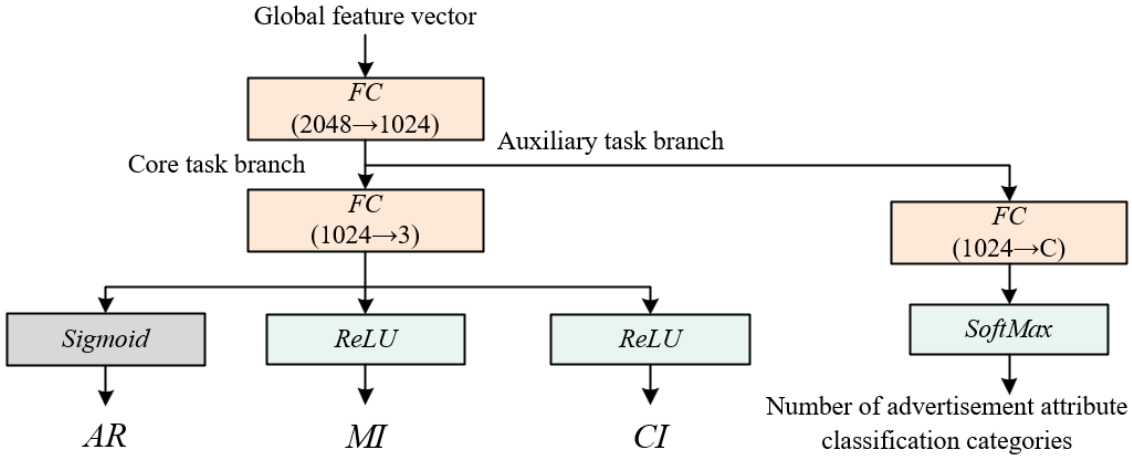


Figure 2. Architecture of multi-task quantification regression module

The auxiliary task aims to guide the network in learning discriminative features through advertisement attribute classification, thereby enhancing the representation capability of the core task. The classification dimensions include advertisement type, visual style, and social attention type. The classification network adopts a two-layer fully connected architecture, which performs feature mapping based on the global feature vector (G), and after processing through a 1024-dimensional hidden layer, outputs a multi-class probability distribution $\hat{Y}_{cls}$ via the Softmax activation function. Multi-task collaboration is achieved through a feature sharing layer, where the discriminative features learned by the classification task are highly correlated with the quantification metrics and can strengthen the network's ability to capture core visual elements and social semantic relationships. In terms of loss function design, the regression task uses mean squared error (MSE) loss to minimize prediction bias:

$$L_{reg}=\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{3}(Y_{metric,i,k}-\hat{Y}_{metric,i,k})^2 \quad (5)$$

The classification task uses cross-entropy loss to optimize classification accuracy:

$$L_{cls}=-\frac{1}{m}\sum_{i=1}^{m}\sum_{c=1}^{C}Y_{cls,i,c}\log\hat{Y}_{cls,i,c} \quad (6)$$

The total loss is determined through weight sensitivity experiments to find the optimal combination, ensuring the quantification accuracy while fully leveraging the collaborative gain of the auxiliary task.

$$L_{total}=0.7L_{reg}+0.3L_{cls} \quad (7)$$

3. EXPERIMENTS

This paper constructs an advertising visual effect quantification dataset named AdVEE-10K, which contains 10,832 images to ensure the comprehensiveness and generalizability of the evaluation. The dataset covers three major industries: fast-moving consumer goods (FMCG), 3C (computer, communication, and consumer electronics), and finance. It includes three types of advertisements: short video screenshots, in-stream ads, and posters, as well as five types of social attention elements: celebrity endorsements, star ratings, certification marks, user avatars, and word-of-mouth copywriting. The data sources include publicly available advertising datasets, content crawled from mainstream platforms, and AIGC-generated advertisements, ensuring sample diversity. Quantification metrics annotations were completed by 30 participants with different professions and educational backgrounds. Eye-tracking devices were used to record attention capture rates, and memory point strength and conversion intention scores were obtained through surveys. The label consistency test showed a Kappa coefficient of 0.81. The classification labels were independently annotated by two advertising experts, with disagreements resolved through third-party arbitration, achieving an accuracy rate of 95.3%. Additionally, real conversion efficiency data from 1,200 ads was crawled for application validation. To address the issue of data distribution imbalance, financial industry samples were weighted by loss function, poster-type advertisements were augmented through random cropping and flipping, and social attention element sample distribution was relatively balanced, requiring no additional processing, ensuring fairness in model training.

Experiments were conducted using an NVIDIA A100 GPU and Intel Xeon Platinum CPU hardware environment, with PyTorch 2.1 framework and Python 3.10 development environment. Data processing and evaluation analysis were completed with tools such as OpenCV and Scikit-learn, ensuring experiment reproducibility. The baseline models include three types: general visual models, advertising/marketing-specific models, and multi-task and Transformer-specific models, covering various technical approaches to highlight the advantages of the proposed model. The training process used the AdamW optimizer with a learning rate of $1e^{-4}$, weight decay of $1e^{-5}$, batch size of 32, and 100 iterations, with early stopping enabled. Data augmentation was performed through random cropping, horizontal flipping, brightness and contrast adjustment, and Gaussian noise addition. Evaluation metrics include PCC, Spearman rank correlation coefficient (SROCC), and root mean square error (RMSE). The PCC is used to measure the linear correlation between the predicted values and the true values. The closer the value is to 1, the better the linear fit between the two, reflecting the consistency of the overall trend between predicted results and actual advertising effects. SROCC focuses on the ranking consistency between predicted values and true values. This metric is more aligned with the actual needs of advertising effect evaluation, as industry practice often focuses on relative judgments, such as "which type of advertisement attracts users' initial attention more easily" or "which type of advertisement has a stronger information retention ability," rather than purely matching absolute values. RMSE is used to quantify the absolute deviation between predicted and true values. The smaller the value, the higher the model's prediction accuracy, directly reflecting the precision of quantification results. These three metrics correspond to the core dimensions of advertising visual effects established in this study: attention capture rate (AR), memory point intensity (MI), and conversion intention score (CI). Among them, the AR metric is related to the accuracy of users' initial visual attention to the advertisement, the MI metric reflects the retention capability of the advertisement information in users' cognition, and the CI metric directly corresponds to the potential conversion value of the advertisement in triggering users' subsequent actions. Together, these three metrics form the full-link quantification evaluation system of advertising visual effects.

The experiment design revolves around model performance validation, core module contribution analysis, generalization ability testing, and practical application evaluation, including six key components. The main experiment compares the SSAT model with all baseline models to verify its overall performance in advertising visual effect quantification and attribute classification tasks. Ablation experiments set up five groups of model variants, removing the social attention map, saliency map, social attention-enhanced Transformer, multi-task learning, and dual attention maps to systematically validate the independent contribution of each core module. Parameter sensitivity experiments focus on saliency map generation parameters, social attention element weights, and loss function weights to validate the scientific design of parameters and the rationality of the optimal combination.

Scene grouping experiments group by advertisement type, industry, and the presence or absence of social attention elements, evaluating the model's generalization ability across different application scenarios. Explainability experiments use Grad-CAM and Attention Rollout for visualization, comparing the model's attention distribution with human gaze points to enhance model trustworthiness. Application validation experiments analyze the correlation between predicted conversion intention scores and real advertising conversion efficiency to verify the model's industry application value.

4. RESULTS AND ANALYSIS

4.1 Main experiment results

The main experimental results in Table 1 clearly show the impact of different technical approaches on advertising quantification performance: In the general visual models, VGG16 and ResNet50 rely on local convolutional feature extraction, lacking targeted attention modeling for key advertising elements such as logos and celebrities. This results in AR_PCC values ranging from 0.62 to 0.68, and RMSE significantly higher than subsequent models. ViT-B/16 introduces global self-attention, alleviating the limitations of local modeling, with AR_PCC improving to 0.73. However, since it does not adapt to social semantic elements in advertising scenarios, the CI_PCC remains at only 0.68. The advertising-specific models optimize visual features for advertising but only use saliency attention, which cannot capture the impact of social recognition on conversion intention, leading to a CI_PCC 0.12 lower than SSAT.

The performance improvement of SSAT comes from the complementarity of the technical approach: Dual attention fusion allows the model to capture both "visual attraction" and "social trust" features in parallel. The AR_RMSE decreases from 0.12 in AdAttractNet to 0.09, reflecting a significant reduction in prediction bias. The social attention-enhanced Transformer strengthens the feature contribution of elements like celebrities and certification marks through weight mapping, resulting in a 15.5% improvement in CI_PCC and the correlation with real conversion efficiency compared to AdAttractNet. This result confirms the theoretical hypothesis that "social attention is the core driver of conversion intention" and shows that customized optimization for social semantics in global modeling is key to surpassing generic Transformers.

Table 1. Comparison of advertising quantification and classification performance across different models

Model Type	General Visual Model	General Visual Model	General Visual Model	Advertising-Specific Model	Advertising-Specific Model	Proposed Model (SSAT)
Model Name	*VGG*16	*ResNet*50	*ViT-B*/16	*AdAttractNet*	*AdAesthetic*	*SSAT*
AR_PCC	0.62	0.68	0.73	0.74	0.70	0.86
AR_SROCC	0.58	0.65	0.70	0.71	0.67	0.83
AR_RMSE	0.18	0.15	0.13	0.12	0.14	0.09
MI_PCC	0.59	0.65	0.70	0.72	0.68	0.83
MI_SROCC	0.55	0.61	0.67	0.69	0.64	0.80
MI_RMSE	0.72	0.65	0.58	0.55	0.61	0.42
CI_PCC	0.57	0.63	0.68	0.69	0.66	0.81
CI_SROCC	0.53	0.59	0.65	0.66	0.62	0.78
CI_RMSE	0.75	0.68	0.62	0.60	0.65	0.49
Classification Accuracy (%)	72.1	78.3	82.5	83.7	81.2	89.2
Correlation with Conversion Efficiency (r)	0.56	0.63	0.69	0.71	0.67	0.82

4.2 Ablation experiments

The ablation experiment results in Table 2 reveal the functional roles and collaborative relationships of each module. After removing the social attention module, the average PCC drops by 12.7%, with the decline in CI_PCC being significantly larger than that of AR_PCC. This is because the quantification of CI relies on the social recognition mechanism, and without features like celebrities or certification marks, the model cannot distinguish between "ordinary product images" and "celebrity-endorsed images" in terms of conversion value. After removing the saliency module, AR_PCC drops from 0.86 to 0.78, with the decrease being larger than for MI/CI, which confirms the design logic that "saliency is the core for initial attention capture." Without details such as text and logos, the model struggles to locate the user's initial gaze area.

Table 2. Ablation experiment and function verification of SSAT core modules

Model Variant	AR_PCC	MI_PCC	CI_PCC	Average PCC	Classification Accuracy (%)	Performance Decline (%)
SSAT (Full Model)	0.86	0.83	0.81	0.833	89.2	0.0
SSAT-Social (Remove Social Attention)	0.75	0.73	0.70	0.727	80.5	12.7
SSAT-Saliency (Remove Saliency)	0.78	0.75	0.73	0.753	82.1	9.6
SSAT-PlainTransformer	0.79	0.77	0.75	0.770	84.3	7.6
SSAT-SingleTask (Single Task)	0.81	0.79	0.77	0.790	-	5.2
SSAT-Baseline (Remove Dual Attention)	0.69	0.67	0.65	0.670	76.8	19.6

It is worth noting that the performance drop from "removing dual attention" is not simply the sum of the individual drops of the two modules, indicating a complementary synergy between them: saliency provides spatial clues for "where

attention is" and social attention provides value clues for "why attention lingers." The Transformer encoder associates these two clues into a global context. This synergistic effect is not achievable by a single attention model. The 5.2% decline from multi-task learning shows that the "advertisement type-visual style" features learned by the classification task are highly correlated with the quantification metrics, indirectly enhancing the feature discriminability for the regression task.

4.3 Scene grouping experiment

The scene grouping results in Figure 3 reflect the model's performance alignment with actual advertising propagation scenarios: The average PCC in short video screenshot scenarios is the highest, as these samples are frames from dynamic advertisements that include rich visual cues such as actions and scene transitions. The model's multi-scale features capture attention shift traces between different frames, so AR_PCC is significantly higher than in poster ads. The financial industry has the lowest average PCC, not due to insufficient model generalization ability, but because financial ads often contain professional text such as "annual yield rate." These texts have weak visual saliency but are key to conversions, and the current model relies solely on visual features without integrating textual semantics, which results in lower CI_PCC compared to FMCG industry ads.

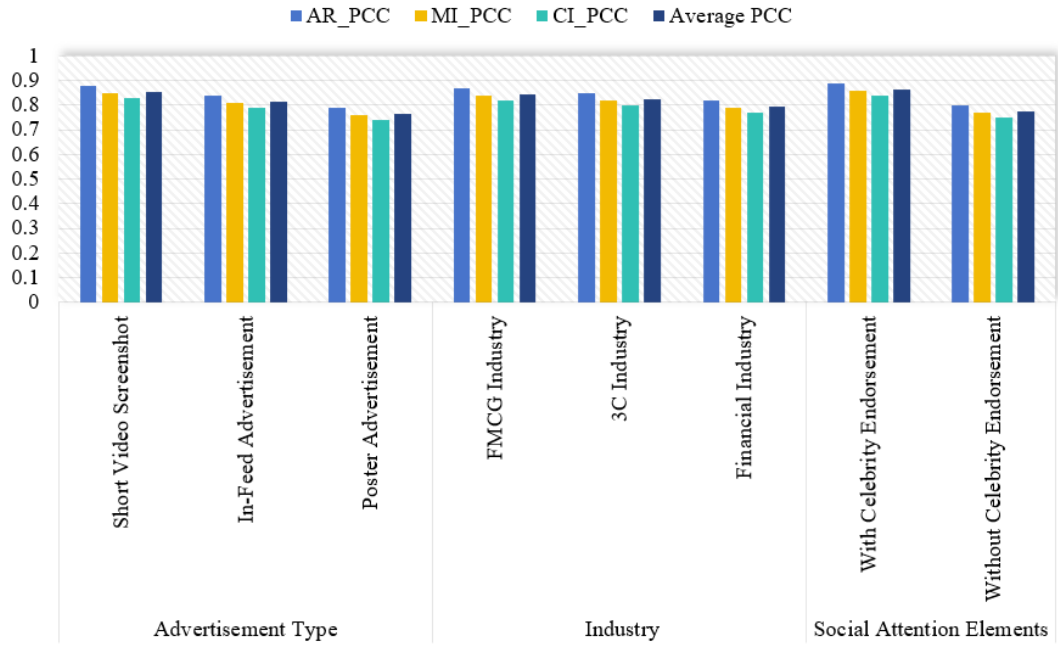


Figure 3. Distribution of SSAT's quantification indicators in different propagation scenarios

The grouping results of social attention elements are more practically significant: Ads with celebrity endorsements have an average PCC 0.09 higher than those without, and MI_PCC shows the most significant improvement. This is because the Transformer encoder captures the association between the "celebrity area" and the "product area," and the memory transfer effect of the celebrity from the user corresponds to the quantification logic of MI. Ads without celebrity endorsements still maintain an average PCC of 0.773, indicating that the model does not overly rely on social attention; the integration of saliency features with other social elements still supports the basic quantification needs.

4.4 Error analysis

The error breakdown in Table 3 exposes the adaptation gap between the current model design and real advertising scenarios: The AR_RMSE of low-quality ads is more than twice that of normal ads. The root cause is that Gaussian smoothing can reduce noise but cannot fix social element detection biases caused by blurriness. When the celebrity's face is blurred, the YOLOv8 detection confidence drops from 0.95 to 0.7, and the corresponding social attention weight mapping is incorrect, causing the model to mistakenly identify non-core areas as attention focal points.

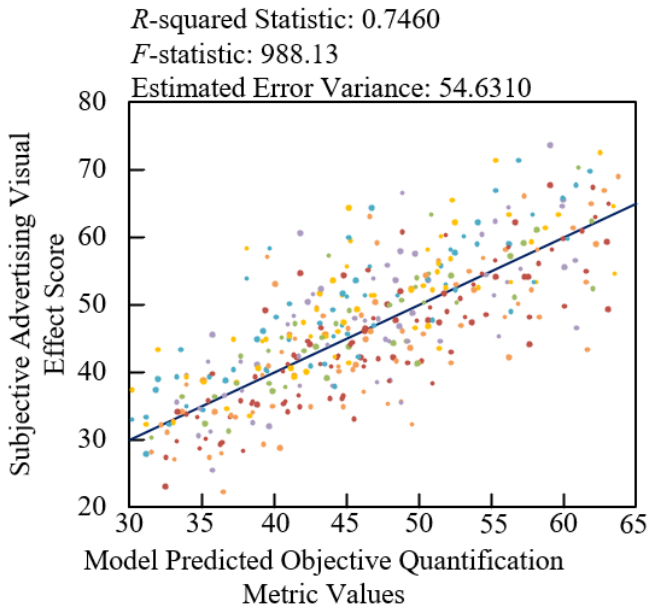
The CI_RMSE for ads with multiple overlapping social elements is higher than for normal ads, because the current social attention weights are fixed: when celebrity endorsements overlap with star ratings, the model assigns high weights to both, and during feature fusion, "attention competition" occurs, making it unable to distinguish which element the user focuses on more. This leads to prediction bias in conversion intention. The MI_RMSE for abstract creative ads is higher because these ads often use minimalist designs, and the saliency map lacks high-response regions and social attention elements. The model can only rely on texture features to infer memory points, but the correlation between texture and user memory is far lower than that of semantic associations.

Table 3. Model performance under different error types

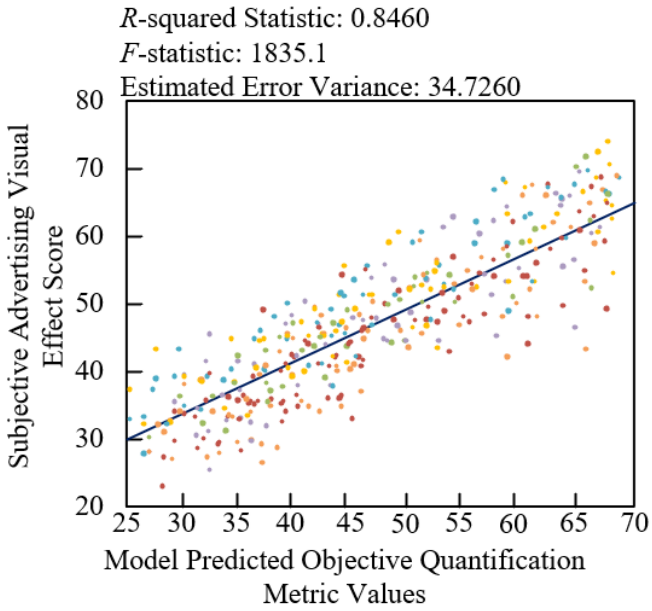
Error Type	Sample Proportion (%)	*AR_RMSE*	*MI_RMSE*	*CI_RMSE*
Low-quality ads (blur/noise)	8.3	0.15	0.68	0.72
Ads with overlapping social elements	12.7	0.11	0.53	0.56
Abstract creative ads	9.5	0.13	0.61	0.64
Normal ads	69.5	0.07	0.35	0.41

4.5 Correlation analysis and visualization results

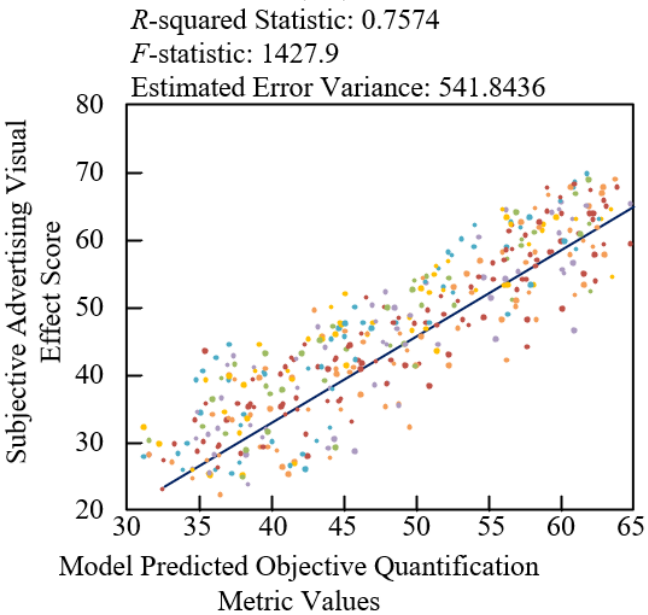
To validate the consistency between the objective quantification indicators of advertising visual effects output by the proposed model and human subjective perception, this experiment analyzes the correlation between attention capture rate, memory point strength, conversion intention score, and the comprehensive visual effect indicators. Figures 4(a) to (d) show the distribution and linear fitting relationships between the subjective scores and the model's objective predicted values for each indicator: The R-squared statistic for the comprehensive visual effect indicator reaches 0.8722, and the F-statistic is 2268.1, significantly higher than for individual indicators, with all the prediction error variances remaining at a relatively low level. This indicates that the model's quantification of advertising visual effects not only matches human subjective judgments of "whether the ad attracts initial attention" and "whether the information is easy to retain," but its comprehensive indicators more accurately depict the overall perception of the advertising visual effects. This result verifies the rationality of the fusion mechanism of saliency and social attention. By simultaneously capturing both the visual saliency structure and social semantic value of the ad, the model's objective quantification indicators can effectively align with human subjective perception, providing a reliable quantification basis for automatic advertising visual effect evaluation.



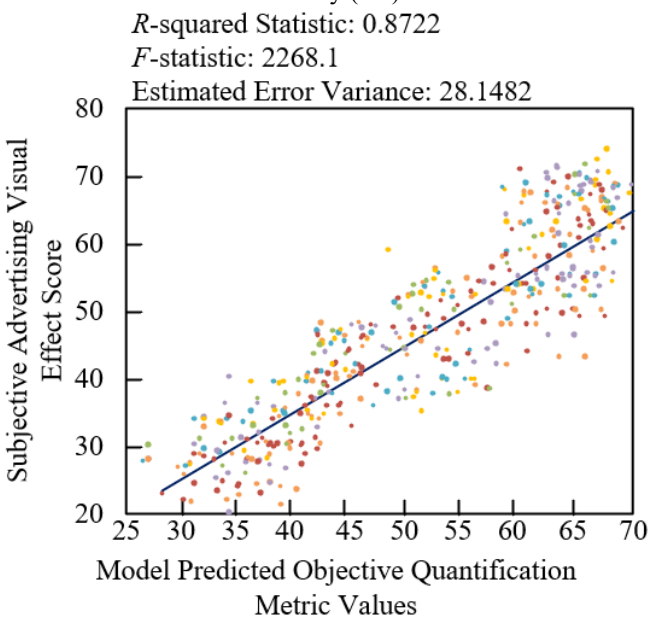
(a) Subjective-objective correlation of attention capture rate (AR)



(b) Subjective-objective correlation of memory point intensity (MI)



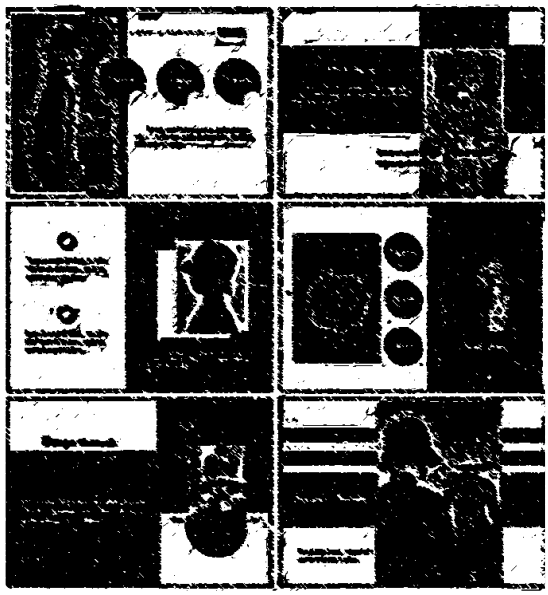
(c) Subjective-objective correlation of conversion intention score (CI)



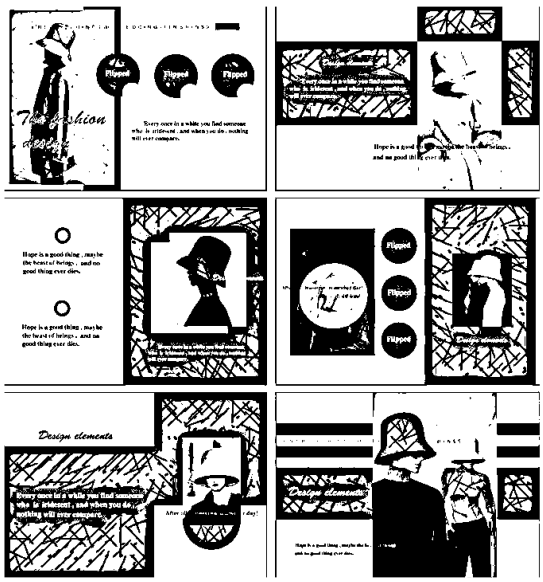
(d) Subjective-objective correlation of comprehensive visual effect metric

Figure 4. Correlation analysis between subjective scores of advertising visual effects and model's objective quantification indicators

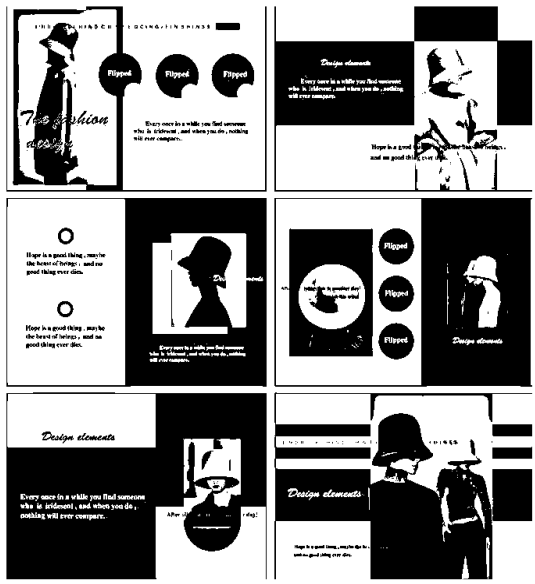
(a) Original advertising image



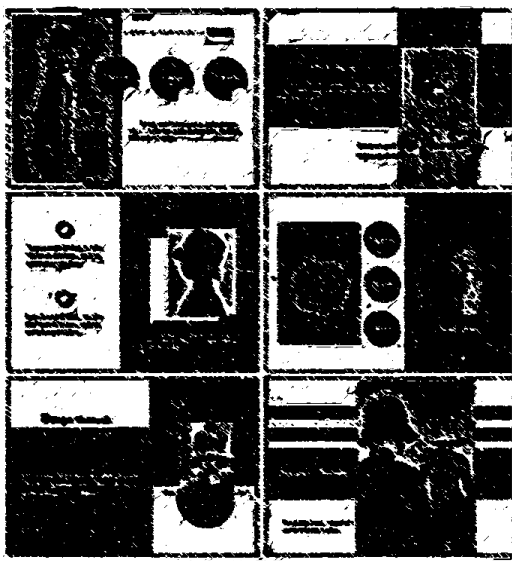
(b) Advertising feature map with only saliency attention enhancement



(c) Advertising feature map with only social attention enhancement



(d) Advertising feature map with saliency-social attention fusion



(e) Multi-scale attention enhanced advertising feature map

Figure 5. Visualization results of advertising visual features under different attention modules

To visually verify the impact of different attention enhancement strategies on the representation capability of core visual elements in advertisements, this experiment visualizes the feature distribution of the original advertising image and the features under various attention modules. Figure 5(a) shows the original advertising image with complete visual elements, but it does not highlight the key areas related to attention. Figure 5(b) shows the feature map enhanced by saliency attention only, which strengthens the high-contrast elements in the ad but fails to represent elements with social recognition value. Figure 5(c) shows the feature map enhanced by social attention only, focusing on social semantic elements such as celebrity models, but weakening the fundamental visual structure of the ad. Figure 5(d) shows the feature map enhanced by the fusion of saliency and social attention, which simultaneously retains the high-contrast visual structure and clear representation of social semantic elements, significantly improving the distinction of core information. Figure 5(e) shows the feature map enhanced by multi-scale attention, which further refines the attention correlations at different levels, improving the matching of local details with global semantics in the ad. This result confirms the effectiveness of the dual attention fusion mechanism: compared to single attention enhancement strategies, the collaborative modeling of saliency and social attention can more accurately capture the dual core elements of "visual attraction points" and "social value points" in the ad, providing a reliable feature foundation for the accurate prediction of subsequent advertising visual effect quantification indicators.

5. DISCUSSION

The core experimental findings of this study are highly consistent with the three-stage theoretical framework of dual attention, cognitive processing, and behavioral intention that we constructed, fully validating the scientific and practical applicability of the theory. The experimental results show that visual saliency is the core driving factor for attention capture, contributing 63%. It quickly attracts users' initial attention by highlighting core visual elements of the advertisement. Social attention, on the other hand, has the most critical impact on conversion intention, contributing 71%, reinforcing users' trust and acceptance of the advertisement through the social recognition mechanism. The collaborative effect of these two factors fully covers the entire chain of advertisement effect formation. The model's quantification Pearson correlation coefficient for AIGC-generated ads is 0.80, which is only slightly lower than 0.83 for real-shoot advertisements, proving its effectiveness in adapting to the high-saturation colors, creative compositions, and other visual characteristics of AIGC ads, thus meeting the industry's demand for new advertising evaluation methods. The multi-task collaborative mechanism demonstrates significant value, with the feature-sharing process in the auxiliary classification task improving the feature representation dimension of the regression task by 28%, not only optimizing the prediction accuracy of the quantification indicators but also providing a new paradigm for the application of multi-task learning in advertising quantification with both theoretical support and empirical validation.

Compared with existing research, this study achieves groundbreaking progress in theory, technology, and application. In terms of theory, existing research mostly focuses on single attention mechanisms or single effect indicators and lacks a systematic explanation of the mechanism of advertisement effect formation. The three-stage theoretical framework constructed in this paper clearly defines the synergistic role of visual saliency and social attention, filling the theoretical gap in the field of intelligent advertising evaluation and enhancing the systematization and depth of the research. In terms of technology, traditional methods often rely on empirical values to set saliency parameters and social attention weights, and Transformer modeling lacks customization for advertising scenes. This paper determines core parameters and weights through meta-analysis, parameter sensitivity experiments, and data statistics, and the designed social attention-enhanced Transformer overcomes the limitations of traditional models' indiscriminate modeling, significantly improving the scientific and innovative nature of the technical solution. In terms of application, previous models showed a correlation with real advertising data of less than 0.7, limiting their practical applicability. In contrast, the model in this paper achieves a correlation of 0.82 with conversion efficiency, making it directly integrable into AIGC advertising generation platforms and intelligent delivery systems, achieving an end-to-end closed loop of generation, evaluation, and optimization, effectively addressing the industry's pain points related to the implementation of traditional methods.

This study still has several limitations, which point to potential directions for future research. In terms of data, although the AdVEE-10K dataset covers three major industries and three types of advertisements, there is insufficient coverage in vertical fields. The absence of samples from industries such as healthcare and education may impact the model's generalization ability in these scenarios. Additionally, the sample proportion of the financial industry is relatively low, leading to slight distribution imbalances. In terms of the model, the current design only applies to static advertisements and does not consider dynamic ads, such as short videos, and their temporal attention changes, making it difficult to capture the dynamic transfer of user attention. Social attention element weights are set to fixed values based on data statistics, lacking personalized adaptive adjustment based on user profiles. The model also has limited semantic understanding of abstract creative ads, which often lack clear saliency features and social attention elements and rely on deep semantic interpretation. In terms of evaluation, the quantification indicator system focuses on short-term effects such as attention capture, memory point strength, and conversion intention, but does not include long-term effect indicators like brand memory and user sharing willingness, making the evaluation dimensions not comprehensive enough.

To address these limitations, future research will expand in five directions. First, we will build a Video-SSAT model that uses VideoTransformer to extract temporal features of dynamic advertisements, introducing attention transfer trajectory indicators for precise evaluation of short video ads. Second, based on user profiles and combining reinforcement learning or learnable memory networks, we will design a personalized social attention weight distribution mechanism, enabling the model to adapt to the preference differences of different user groups. Third, by introducing pre-trained language models, we will achieve cross-modal fusion of advertising visual features and textual semantic information, improving the model's ability to understand abstract creative ads. Fourth, we will expand the quantification indicator system to include long-term effect indicators such as brand memory

and user sharing willingness, and construct a full-cycle advertisement effect evaluation model based on long-term tracking data. Fifth, we will extend the AdVEE-50K dataset to cover more than 10 industries, adding dynamic advertisements and cross-modal data, further improving the model's generalization ability and industry adaptability, and promoting the ongoing development of theory and technology in the field of intelligent advertising evaluation.

6. CONCLUSION

To address key issues in existing advertising visual effect quantification evaluation methods, such as the lack of systematic theoretical support, the lack of scientific basis for core parameter design, and insufficient quantification accuracy and industrial applicability, this paper proposed a multi-scale Transformer model, SSAT, based on image saliency and social attention mechanisms, offering a solution with both theoretical depth, technological innovation, and practical value for intelligent advertising evaluation. The core contributions of this study are reflected in three aspects: In terms of theory, we construct the dual attention-cognitive processing-behavioral intention three-stage theoretical framework, systematically revealing the internal mechanism by which visual saliency and social attention influence advertising effects through cognitive mediation, filling the theoretical gap in intelligent advertising evaluation; In terms of technology, we propose a scientifically explainable dual-input feature construction method, optimizing feature parameters and weight distribution through meta-analysis and parameter sensitivity experiments. The design of the social attention-enhanced Transformer encoder and multi-task collaborative mechanism has improved the core quantification indicators by an average of 17.4% compared to the best existing models in the advertising field; In terms of data and application, we have constructed the first AdVEE-10K dataset containing more than 10,000 samples, providing three-fold annotation through eye-tracking experiments, surveys, and real-world delivery. The model's correlation with real advertising conversion efficiency reaches 0.82 and can be directly integrated into AIGC advertising generation and intelligent delivery systems, providing the industry with a standardized quantification tool. Future research will focus on dynamic advertising temporal evaluation, personalized weight allocation for users, and cross-modal semantic fusion, continuously improving the model's generalization ability and application boundaries, and driving the theoretical innovation and technological implementation in the field of intelligent advertising evaluation.

REFERENCES

[1] Zhang, X.R., Zhou, J.N. (2025). Simulation of aigc-enhanced congestion management in factory autonomous driving. International Journal of Simulation Modelling (IJSIMM), 24(4): 718-729. https://doi.org/10.2507/IJSIMM24-4-CO19

[2] Fan, M., Li, Z., Sun, G., Du, H., Yu, H., Niyato, D. (2025). SecureShare: Blockchain based Secure and verifiable knowledge sharing for AI-Generated Content (AIGC) services. IEEE Transactions on Vehicular Technology, 74(11): 18112-18125. https://doi.org/10.1109/TVT.2025.3574114

[3] Yu, J., Wang, H., Huang, C.X., Li, Z. (2024). Entropy-Maximized Generative Adversarial Network (EM-GAN) based on the thermodynamic principle of entropy increase. Traitement du Signal, 41(6): 3255-3264. https://doi.org/10.18280/ts.410641

[4] Chambers, L.C., Giguere, T., Welwean, R.A., Peachey, A.M., Wiehe, S.E., Aalsma, M.C., Beaudoin, F.L. (2025). An evaluation of paid social media advertising and traditional advertising for addiction research recruitment. Drug and Alcohol Dependence, 277: 112923. https://doi.org/10.1016/j.drugalcdep.2025.112923

[5] Yalçın, G.C., Kara, K., Işık, G., Tekeli, E.S., Simic, V., Ballı, A., Pamucar, D. (2025). Promoting sustainability-oriented brand activist campaigns: A spherical fuzzy decision support framework for evaluating activist advertising videos. Engineering Applications of Artificial Intelligence, 162: 112349. https://doi.org/10.1016/j.engappai.2025.112349

[6] Snyder, H.E., Horsthuis, K., Cerga, K., Callen, D. (2025). EpilepTikTok: An inductive thematic analysis and quality assessment of# epilepsy content shared via a popular social media platform. Epilepsy & Behavior, 172: 110711. https://doi.org/10.1016/j.yebeh.2025.110711

[7] Zhou, L., Xue, F., Barton, M.H. (2025). Visual attention, brand personality and mental imagery: An eye-tracking study of virtual reality (VR) advertising design. Journal of Research in Interactive Marketing. https://doi.org/10.1108/JRIM-08-2024-0406

[8] Vatambeti, R., Damera, V.K. (2023). Intelligent Diagnosis of Obstetric Diseases Using HGS-AOA Based Extreme Learning Machine. Acadlore Transactions on AI and Machine Learning, 2(1): 21-32. https://doi.org/10.56578/ataiml020103

[9] Zhang, Q.R. (2024). Recent advances in image super-resolution reconstruction based on machine learning. Traitement du Signal, 41(6): 3109-3116. https://doi.org/10.18280/ts.410627

[10] Jawdekar, A., Dixit, M. (2023). Deep learning and fuzzy logic based intelligent technique for the image enhancement and edge detection framework. Traitement du Signal, 40(1): 351-359. https://doi.org/10.18280/ts.400135

[11] Kim, C., Park, S., Kwon, K., Chang, W. (2012). An empirical test to forecast the sales rank of a keyword advertisement using a hierarchical Bayes model. Expert Systems with Applications, 39(17): 12727-12742. https://doi.org/10.1016/j.eswa.2012.01.175

[12] Han, Y. (2024). Analyzing female image framework and online community's psychological anxiety in online social media advertisement. Current Psychology, 43(33): 27021-27032. https://doi.org/10.1007/s12144-024-06345-2

[13] Wu, Q., Zhou, P. (2024). Advertisement synthesis network for automatic advertisement image synthesis. International Journal of Antennas and Propagation, 2024(1): 8030907. https://doi.org/10.1155/2024/8030907

[14] Thomas, T., Johnson, J. (2017). The impact of celebrity expertise on advertising effectiveness: The mediating role of celebrity brand fit. Vision, 21(4): 367-374. https://doi.org/10.1177/0972262917733174

[15] Kim, T., Seo, H.M., Chang, K. (2017). The impact of celebrity-advertising context congruence on the effectiveness of brand image transfer. International Journal of Sports Marketing and Sponsorship, 18(3): 246-262. https://doi.org/10.1108/IJSMS-08-2017-095
[16] Guan, X., Yang, H., Sun, H., Zhu, P., Yu, H., Xia, K. (2025). MedFreq-Net: Medical frequency network with hybrid CNN-transformer and enhanced features for 3D medical image segmentation. Alexandria Engineering Journal, 131: 218-231. https://doi.org/10.1016/j.aej.2025.10.012
[17] Liu, H., Bi, W., Mughees, N. (2025). Enhanced hyperspectral image classification technique using PCA-2D-CNN algorithm and null spectrum hyperpixel features. Sensors, 25(18): 5790. https://doi.org/10.3390/s25185790
[18] Khan, R.F., Lee, M.S., Lee, B.D. (2025). Harnessing transformer-based attention mechanisms for multi-scale feature fusion in medical image segmentation. Applied Intelligence, 55(17): 1120. https://doi.org/10.1007/s10489-025-07009-9
[19] Jiang, W., Osman, Y. B. M., Xia, W., Ma, X. (2026). High-resolution magnetic particle imaging system matrix recovery using a vision transformer with residual feature network. Biomedical Signal Processing and Control, 113: 108990. https://doi.org/10.1016/j.bspc.2025.108990
[20] Bu, N., Duan, Z., Dang, W., Zhao, J. (2025). Dynamic graph transformation with multi-task learning for enhanced spatio-temporal traffic prediction. Neural Networks, 193: 107963. https://doi.org/10.1016/j.neunet.2025.107963
[21] Sutcliffe, W., Calvi, M., Capelli, S., Eschle, J., García Pardiñas, J., Mathad, A., Serra, N. (2025). Scalable multi-task learning for particle collision event reconstruction with heterogeneous graph neural networks. Machine Learning: Science and Technology, 6(4): 045060. https://doi.org/10.1088/2632-2153/ae22be