Future Technology
Open Access Journal
ISSN 2832-0379

February 2027| Volume 06 | Issue 01 | Pages 14-23

Journal homepage: https://fupubco.com/futech

https://doi.org/10.55670/fpll.futech.6.1.2

Future Publishing LLC

Article

A multimodal deep learning model integrating visual features and early comment sentiment for predicting audience acceptance of Tibetan-motif tourism advertisements on Xiaohongshu

Lu Hong, Mohd Nor Shahizan Ali*, Mohd Azul Mohamad Salleh

Fakulti Sains Sosial dan Kemanusiaan, Universiti Kebangsaan Malaysia, Bangi 43600, Malaysia

ARTICLE INFO

Article history:
Received 15 June 2026
Received in revised form
10 October 2026
Accepted 03 November 2026

Keywords:
Multimodal deep learning,
Cross-modal attention fusion,
Comment sentiment analysis,
Tibetan ornamental motifs,
Tourism advertising acceptance

*Corresponding author
Email address:
shahizan@ukm.edu.my

DOI: 10.55670/fpll.futech.6.1.2

ABSTRACT

Tibetan art's ornate designs are widely used in social media promotions in the tourism industry, but predicting consumer reactions to ethnic advertising remains difficult. For some people, the elaborate ornamentation symbolizes cultural authenticity; for others, it seems staged. Most existing models analyze imagery and user commentary in isolation. Whether these two sources offer complementary predictors of audience response therefore remains unclear. Here we present a gated bidirectional cross-attention model to predict thirty-day audience acceptance of Tibetan-motif tourism advertisements on Xiaohongshu from visual representations and early twenty-four-hour comment sentiment. We first train the motif-recognition backbone and two interpretable descriptors (motif density and circular hue chromatic conformity) on 19,142 archived pattern images, then transfer them to 4,860 promotional posts using a month-stratified partitioning scheme. Acceptance is treated as a three-category ordinal target derived from engagement residuals adjusted for covariates. Sentiment measures calculated from 149,300 preliminary comments feed into an entropy-dependent adaptive gate that shifts the balance toward visual information as discussion threads become scarce. Our proposed model achieves 0.754 accuracy and 0.734 macro-F1 in experiments on hand-crafted, unimodal, concatenation, CLIP, and ViLT baselines. It beats standard late fusion by 3.5 and 3.7 percentage points on these two metrics and cuts ordinal Mean Absolute Error (MAE) by 12.3 percent relative (from 0.318 to 0.279). Cross-modal complementarity is validated by Accumulated Local Effects (ALE) and Ablation Analyses, with chromatic conformity and sentiment polarity increasing acceptance, while motif density exhibits an inverse-U relationship.

1. Introduction

1.1 Research background and problem statement

Conventional Tibetan ornaments hold tremendous symbolic significance within Chinese ethnic material culture. Whether in woven fabrics, temple timber, or thangka paintings, geometric grids, auspicious Buddhist symbols, and the rigorous use of mineral colors have long conveyed cosmological, religious, and communal values. These ornaments now appear prominently in various online tourism promotions. As platform-based destination marketing grows, algorithms, user behavior, and capital work together to reorganize the economy of tourist destinations, both symbolically and geographically [1]. Visual uniqueness is what makes ethnic-cultural destinations stand out as market differentiation tools. In this regard, Little Red Book (Xiaohongshu) is a visual discovery application with a comment box that collects user feedback and shapes others' future perceptions of a post. Research conducted on User Generated Content (UGC) shows that travel recommendations posted on Xiaohongshu influence travel decisions much more strongly than recommendations posted on social media sites such as Weibo [2]. This is not due to the larger audience. Travelers visiting these sites analyze the visual message from the comments' perspective and assess the credibility of the culture. Destination Marketing Organizations (DMOs) still lack a way to estimate in advance the success of culturally oriented creative materials. Ornament that shows the authenticity of heritage for some audiences can appear as commercialization to others. Traditional focus group studies are not equivalent here

because they rely on self-reporting in laboratory conditions rather than actual mobile scrolling experience.

Abbreviations

ALE	Accumulated Local Effects
AUC	Area Under the ROC Curve
BERT	Bidirectional Encoder Representations
CLIP	Contrastive Language-Image Pre-training
CNN	Convolutional Neural Network
CORAL	Consistent Rank Logits
CV	Cross-Validation
DMO	Destination Marketing Organization
GAM	Generalized Additive Model
HS	Hue-Saturation
MAE	Mean Absolute Error
MHA	Multi-Head Attention
PCA	Principal Component Analysis
PIPL	Personal Information Protection Law
RoBERTa	Robustly Optimized BERT Approach
UGC	User-Generated Content
ViLT	Vision-and-Language Transformer
ViT	Vision Transformer

1.2 Related work and research gap

Four related fields shape our research: computational heritage recognition, sentiment analysis from social media, multimodal engagement prediction, and ordinal classification. Deep Residual Learning (ResNet) in computer vision addressed gradient vanishing in deep architectures and enabled the extraction of fine-grained hierarchical textures [3]. Deep convolutional filters are crucial for identifying patterns in woven or painted folk ornaments because class boundaries rely on repetitive local patterns rather than broad color characteristics. Afterward, Vision Transformers (ViT) revitalized image recognition through sequence modeling based on flattened image patches, where self-attention layers capture global compositional relationships within a single design [4]. Regarding Tibetan arts, CNNs and attention mechanisms have been used for pattern recognition and thangka segmentation [5,6]. Even so, those works judge models strictly against expert category labels. They never link visual descriptors to audience response.

In text modeling, Masked Language Pre-training via BERT and RoBERTa models were used to generate transferable sentence representations [7,8]. Emotion is rarely encoded within colloquial, short Chinese social media messages with single words. Instead, the affective component is expressed through modal particles, reduplication, and discourse context. Hence, domain-specific bidirectional encoders outperform lexicon rules and recurrent neural network-based models for Chinese hospitality and tourism reviews [9]. In the tourism communication literature, evidence suggests that social affordances, parasocial interaction, and endorser symbol congruence influence the creation of destination images [10-12]. Yet, empirical studies on tourism continue to separate visual and textual elements.

In multimodal marketing analytics, computer vision techniques have been employed for forecasting Instagram engagement in food marketing [13], whereas recent research in advertising has indicated that consumer perceptions of digital and generative advertising depend on appeal type and the interaction of the visual and verbal [14-16]. Reviews of multimodal sentiment analysis [17] and multi-task cross-attention architectures [18, 19] also confirm that cross-modal fusion outperforms unimodal architectures, and vision-language models such as Contrastive Language-Image Pre-training (CLIP) [20] and Vision-and-Language Transformer (ViLT) [21] perform well on the problem of caption-image alignment. There remain three important research gaps. First, heritage computing literature categorizes pattern classes [5, 6] without quantifying how the interpretability of cultural characteristics (motif density and chromatic consistency) affects user acceptance. Second, promotional pictures show a weaker association with discussion posts than caption-image pairs [20,21]; moreover, older posts receive lower discussion rates, degrading uncalibrated fused neural networks. Third, previous popularity predictors [22] use lifetime discussion statistics as the target label and extract features from discussions. This approach creates label-input circularity and overlooks the ordinal nature of hierarchical acceptance processes estimated by the Consistent Rank Logits (CORAL) [23] algorithm.

1.3 Research questions, contributions, and organization

Our study rests on a direct hypothesis: visual cultural arrangement and early audience sentiment provide partially independent, complementary signals of long-term advertisement acceptance. We tested three research questions. (RQ1) Does integrating visual representations with early twenty-four-hour comment sentiment via an entropy-aware gated cross-attention network significantly beat handcrafted, unimodal, and vision-language baselines (CLIP, ViLT, CORAL) in predicting thirty-day ordinal acceptance? (RQ2) How well does stochastic modality dropout paired with adaptive gating preserve predictive accuracy when promotional posts have sparse or zero early comments (cold-start/pre-launch scenarios)? (RQ3) How strongly do interpretable cultural descriptors (motif density and circular-hue chromatic conformity) drive acceptance, and are their marginal effects monotonic?

This study makes three primary contributions. Methodologically, we build an ordinal-aware gated bidirectional cross-attention architecture that dynamically weights visual and sentiment streams by thread length and entropy; restricting text inputs to an early twenty-four-hour comment window eliminates temporal circularity and supports pre-launch visual screening. Empirically, we formulate and expert-validate two culturally grounded descriptors (attention-weighted motif density and circular-hue Wasserstein chromatic conformity) through a two-stage transfer scheme. For management, we apply Accumulated Local Effects (ALE) [24] without accounting for metadata to show an inverted-U relationship between motif density and a significant positive effect of chromatic harmony, which supports respectful cultural design and automated creative filtering. Section 2 discusses the specifics of our framework, dataset, descriptor, and fusion method. Section 3 provides benchmarking results, ablation studies, and attributions, while Sections 4 and 5 focus on the discussion and conclusion.

2. Materials and methods

2.1 Research framework and overall procedure

Our modeling workflow operates in four distinct phases, as illustrated in Figure 1. As shown in Figure 1, Phase I obtains source-domain supervision from public pattern repositories and Chinese sentiment datasets. Phase II obtains dual-modal visual features together with culturally interpretable attributes (motif density ρ and chromatic conformity κ) along with early twenty-four-hour thread sentiment (polarity s, intensity ι, and dispersion δ). In Stage III, an adaptive

sparsity gate regulates bidirectional cross-modal attention. Stage IV then optimizes an ordinal focal and modality-consistency objective to predict thirty-day stratified acceptance tertiles. It is crucial to distinguish between capability acquisition (Stage I) and task learning (Stages II-IV) in our approach. While curated pattern archives provide expert-vetted taxonomic guidance, they lack engagement data, whereas promotional posts posted by the platform provide behavioral information but are unannotated at the pixel level. Keeping the core feature extractors frozen after Stage I helps maintain the culture-driven interpretation of ρ and κ without letting them become some engagement-dependent but opaque counterparts. We keep modality-specific encoders in separate multi-head self-attention layers [25] before cross-modal interaction, to avoid being overwhelmed by high-dimensional visual features while processing low-dimensional sentiment. We also use stochastic modality dropout (0.15) during training. Each branch thus remains informative on its own for pre-launch or cold-start posts.

2.2 Dataset construction, preprocessing, and temporal protocol

Table 1 provides an overview of our datasets, where source-domain modules are trained using domain-specific data, while the target-domain tasks are recorded independently. For Stage I vision transfer, 6,842 validated Tibetan pattern samples belonging to eight classical classes (Endless Knot, Lotus, Cloud-and-Thunder, Eight Auspicious Symbols, Geometric Fretwork, Snow Lion, Longevity Crane, and Mandala Borders) were used along with 12,300 additional traditional patterns (for example, Han brocade, Miao embroidery) as negatives so that the classifier doesn't confuse general pattern density with Tibetanness. For Stage I text transfer, we use an initialized Chinese RoBERTa-wwm-ext (12 layers, 768 hidden units) encoder [7,8] pre-trained for two epochs by continuing masked-language modeling on 50,000 unlabeled Xiaohongshu travel posts and further fine-tune on 128,000 labeled reviews from ChnSentiCorp (hotel texts: 10,000), Dianping (lifestyle reviews: 58,000) and Weibo-Senti100k (social posts: 60,000).

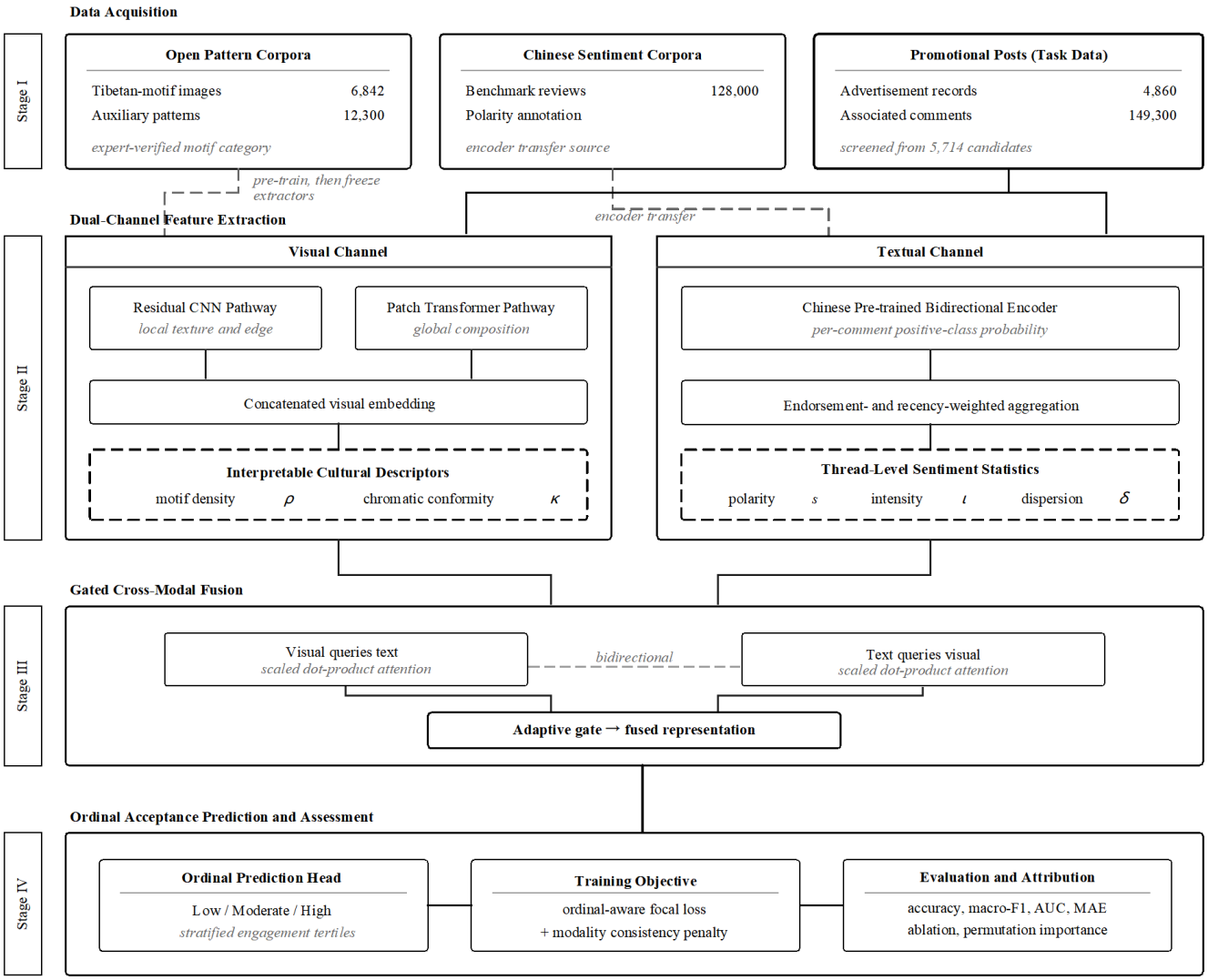


Figure 1. Overall research framework and four-stage pipeline covering Stage I data collection, Stage II dual-channel feature extraction, Stage III gated cross-modal fusion, and Stage IV ordinal acceptance prediction

Table 1. Dataset composition, functional roles, sample scales, and annotation sources

Component	Role	Scale	Annotation Protocol
Tibetan-motif images	Module training	6,842	Expert-verified 8-class taxonomy
Auxiliary pattern images	Contrastive training	12,300	Binary non-Tibetan pattern class
Chinese sentiment corpora	Encoder transfer	128,000	Verified polarity labels
Promotional posts	Task data (t = 30d)	4,860	Covariate-adjusted ordinal tertiles
Early-window comments	Task data (t <= 24h)	149,300	Transferred probabilistic scores

As far as target task data was concerned, we collected public advertisements on Xiaohongshu (2022-2024) following three criteria: (1) verified Tibetan imagery recognized by the classifier of the Stage I (confidence >= 0.80), (2) promotion intention evidenced through destination tags, trip links, or labeling of commercial nature, and (3) at least three comments posted in the first 24 hours after posting. Among the total of 5,714 potential ads, 512 ads with 0 to 2 early comments were separated into a cold-start/pre-launch test dataset, and 226 were excluded because of heavy occlusion, while 116 were excluded due to their non-promotional travelogue nature. This left 4,860 primary analyzable posts and 149,300 early comments.

The circularity that stems from label input is evident when comment counts are fed into the target index, and sentiments come from lifecycle comments. This circularity is avoided in our analysis as we confine the input statistics of sentiments exclusively to the comments collected until the first 24 hours (t<=24h), while measuring the true acceptance on day 30 (t=30d) using only cumulative likes (L_j), saves (F_j), and shares (S_j) but not comments (C_j). The raw engagement is also affected by the creator's reach and any paid promotion. We therefore use log-transformed measurements on day 30 and then residualize them by ordinary least squares regression on five covariates: log number of followers, institutional account dummy, sponsored post dummy, caption length, and posting day of week. Following distributional aesthetic evaluation [26], we compute the composite engagement index E_j for post j from the first Principal Component Analysis (PCA) loading of the standardized covariate-adjusted interaction matrix in Eq (1).

$$E_j = \eta_1 L_j + \eta_2 F_j + \eta_3 S_j \tag{1}$$

where L̃_j, F̃_j, and S̃_j denote the covariate-adjusted log-engagement residuals for likes, saves, and shares of post j, weighted by first principal component loadings η_1 = 0.84, η_2 = 0.61, and η_3 = 0.47 (76.4% of variance). In a sensitivity check that retains 30-day comment residuals C̃_j (η_4 = 0.39), the four-variable index shows a 0.964 Spearman rank correlation with Eq. (1) and shifts downstream classification accuracy by under 0.2 percentage points. We standardize E_j within 288 strata jointly defined by posting month (36 months) and primary motif class (8 classes), and use the training-set tertiles of these stratified residuals (Z_j <= -0.431, -0.431 < Z_j <= 0.438, Z_j > 0.438) to define Low (1), Acceptance scores are categorized into Moderate (2) and High (3). Tertiles provide unambiguous triage layers for marketers and mitigate continuous ratings [26].

The images are resized to 384 × 384 pixels using reflection padding to maintain the periodic nature of motifs, and template-based bot responses are eliminated through MinHash duplicate detection.

2.3 Multimodal feature extraction and fusion model

Figure 2 illustrates the diagram of the multimodal feature extractor. The visual encoder pairs a local convolutional pathway (ResNet-50 [3]) with a global patch-sequence pathway (ViT-B/16 [4] with N_p = 576 patches of size 16 x 16); Figure 2 also shows the domain-adapted Chinese RoBERTa sentiment stream, the interpretable descriptors, the bidirectional cross-attention module, and the adaptive modality gate.

In Stage I, ResNet-50 and ViT-B/16 are trained on the 19,142 pattern images for 9-way classification (8 Tibetan motif classes plus 1 contrastive class) and patch-level motif detection. On a 10% held-out source split, they reach 94.2% top-1 accuracy. When transferring to Stage II promotional photographs, we freeze stages conv1–layer3 of ResNet-50, blocks 1–9 of ViT-B/16, and the linear patch motif probe w_m, and fine-tune only layer4 and blocks 10–12. Motif density ρ_j is then defined as the attention-weighted fraction of patches containing verified ornamental content in Eq (2).

$$\rho_j = \frac{\sum_{p=1}^{N_p} \alpha_{j,p}\sigma(w_m^{\mathrm{T}} z_{j,p})}{\sum_{p=1}^{N_p} \alpha_{j,p}} \tag{2}$$

where α_{j,p} is the multi-head attention rollout weight assigned to patch p of post j, and σ(w_m^T z_{j,p}) is the frozen Stage I patch motif probability. Our second descriptor, chromatic conformity κ_j, measures alignment with canonical Tibetan mineral palettes (vermilion, ultramarine, mineral green, orpiment yellow, and white). Three Tibetan art experts at Southwest Minzu University curated a reference 2D Hue-Saturation (HS) histogram H_ref (36 x 10 = 360 bins) from 500 heritage artifacts. Standard Euclidean distances distort red-boundary transitions because hue h in [0°, 360°) is circular. We therefore define the circular ground distance d_circ(u, v) between bins u = (h_u, s_u) and v = (h_v, s_v) in Eq (3) and the normalized first-order Wasserstein chromatic conformity κ_j in Eq (4).

$$d_{\mathrm{circ}}(\mathrm{u},\mathrm{v}) = \sqrt{\left(\frac{\min(|h_u - h_v|, 360° - |h_u - h_v|)}{180°}\right)^2 + (s_u - s_v)^2} \tag{3}$$

$$\kappa_j = 1 - \frac{W_1(H_j, H_{\mathrm{ref}})}{\max_{H'} W_1(\mathrm{H}', H_{\mathrm{ref}})} \tag{4}$$

Moving from flat patterns to images changes lighting and angles. To confirm both metrics, we had three annotators mark 300 Xiaohongshu post images for motif type, subjective decoration density (Krippendorff's α = 0.84), and color consistency (α = 0.81). In these 300 images, the Stage I classifier demonstrated motif precision at 86.7%, while ρ_j agreed well with the annotators' rating of the density (r = 0.82, p < 0.001, against r = 0.61 for non-weighted class-token attention) and κ_j had high correlation with color ratings (r = 0.79, p < 0.001, against r = 0.68 for Euclidean Wasserstein distance). How well does our text encoder handle platform slang? On a manually annotated target sample of 500 Xiaohongshu tourism comments (Fleiss' κ = 0.83), the domain-adapted Chinese RoBERTa encoder reached 88.4% polarity accuracy and 0.876 macro-F1 [9].

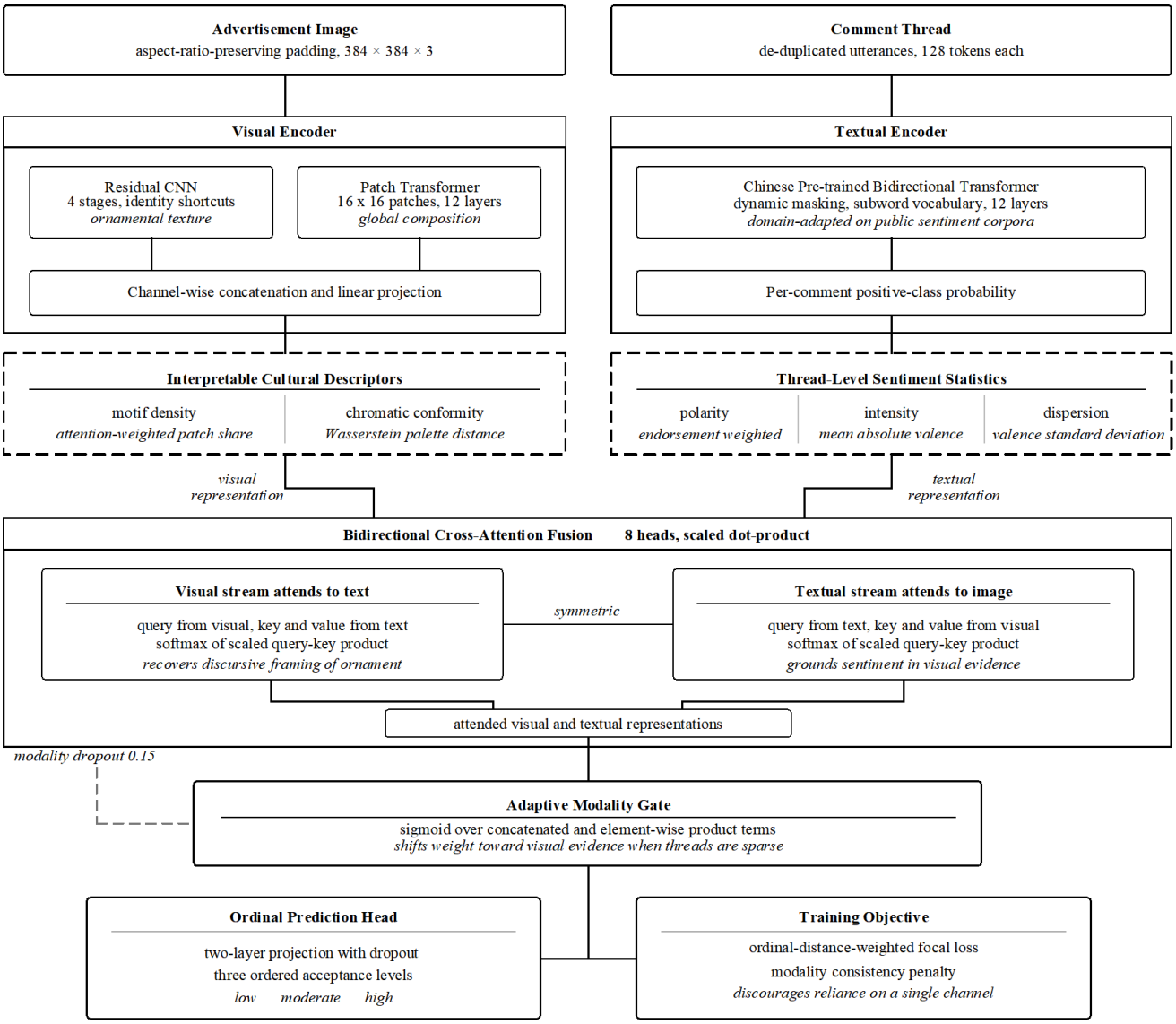


Figure 2. Architecture of the multimodal feature extractor and cross-attention fusion network showing dual visual pathways, sentiment aggregation, bidirectional cross-attention, and the adaptive modality gate

From per-comment positive-class probabilities p_{j,k}^+ across M_j early comments (t <= 24 h), we map each comment to a valence score v_{j,k} = 2 p_{j,k}^+ - 1 in [-1, 1] and aggregate thread polarity s_j using weights ω_{j,k} for peer endorsement (comment likes l_{j,k}) and recency Δt_{j,k} (decay constant λ = 0.05 day^-1), as in Eq (5). Companion metrics for sentiment intensity ι_j and dispersion δ_j are given in Eq (6).

$$s_j = \frac{\sum_{k=1}^{M_j} \omega_{j,k}(2p_{j,k}^{+}-1)}{\sum_{k=1}^{M_j} \omega_{j,k}}, \omega_{j,k} = \ln(1 + l_{j,k})e^{-\lambda \Delta t_{j,k}} \quad (5)$$

$$\iota_j = \frac{1}{M_j}\sum_{k=1}^{M_j} |2p_{j,k}^{+} - 1|, \delta_j = \sqrt{\frac{1}{M_j}\sum_{k=1}^{M_j}(v_{j,k} - \bar{v}_j)^2} \quad (6)$$

where v̄_j is the simple average valence; δ_j helps distinguish between polarized threads (combination of approval and objections) and indifference threads. Two-way cross-attention [18] allows visual tokens to ask questions to textual tokens, and vice versa, generating attended vectors ṽ_j and t̃_j (alongside ρ_j, κ_j and s_j, ι_j, δ_j). Promotional images and their comment threads are not as strongly coupled as image captions, and the number of comments at early stages can be highly variable. An adaptive gate g_j in (0, 1)^d conditioned on normalized log thread length ñ_j = ln(1+M_j)/ln(1+M_max), mean comment Shannon entropy S̄_j, and dispersion δ_j therefore regulates modality fusion in Eqs (7) and (8).

$$g_j = \sigma(W_g[\tilde{v}_j; \tilde{t}_j; \tilde{v}_j \odot \tilde{t}_j; \tilde{n}_j; S_j; \delta_j] + b_g + \tau e^{-\gamma M_j} 1) \quad (7)$$

$$z_{f,j} = g_j \odot \tilde{v}_j + (1 - g_j) \odot \tilde{t}_j \quad (8)$$

In Eq. (7), τ e^{-γ M_j} (with learnable τ initialized to 1.5 and γ = 0.25) acts as an explicit sparsity prior. For sparse threads (M_j <= 3), mean gate weight ḡ_j averages 0.88 ± 0.04 so predictions rely mainly on visual evidence; at median length (M_j = 19), ḡ_j balances at 0.54 ± 0.06; and for dense threads (M_j >= 50), ḡ_j shifts to 0.39 ± 0.05. Network weights are trained by minimizing an ordinal-aware focal loss

combined with a Kullback-Leibler (D_KL) consistency penalty against modality-masked predictions ŷ_j^{(-m)} in Eq (9).

$$L = -\sum_{c=1}^{3} \phi(c, y_j)(1-\hat{p}_{j,c})^{\gamma} \ln \hat{p}_{j,c} + \beta \sum_{m \in \{v,t\}} D_{KL}(\hat{\boldsymbol{y}}_j \parallel \hat{\boldsymbol{y}}_j^{(-m)}) \quad (9)$$

where ϕ(c, y_j) = 1 + |c - y_j| grows with ordinal distance and ŷ_j^{(-m)} denotes the prediction when modality m is replaced by a learned zero-information mask token m_0. We evaluate feature attribution across all six descriptors (chromatic conformity κ, sentiment polarity s, compositional embedding, motif density ρ, sentiment intensity ι, and thread dispersion δ) via permutation importance Δ_f over R = 20 shuffles and Accumulated Local Effects (ALE) [24]. Ordinal error is measured by ordinal Mean Absolute Error (MAE_ord) in Eq (10).

$$\Delta_f = \mathrm{F1}^{\mathrm{full}} - \frac{1}{R}\sum_{r=1}^{R} \mathrm{F1}_f^{(r)}, \mathrm{MAE}_{\mathrm{ord}} = \frac{1}{N_{\mathrm{test}}}\sum_{j=1}^{N_{\mathrm{test}}} |\hat{y}_j - y_j| \quad (10)$$

3. Results

3.1 Experimental setup and evaluation metrics

The 4,860 posts that can be analyzed will be divided into training (25 months, n=3,402, 70%), validation (4 months, n=486, 10%), and testing (7 months, n=972, 20%) datasets. Clustering by month of publication avoids temporal leakage, whereas stratification by acceptance tertile and motif class produces balanced splits: Training (Low: 33.1%, Moderate: 33.5%, High: 33.4%; imbalance = 1.01), Validation (32.9%, 33.7%, 33.3%), and Testing (33.2%, 33.1%, 33.6%). We also run a 5x2 repeated CV across monthly groups to test split variance beyond the seed-initialization stage. All baselines have the same budget for hyperparameter tuning (20 Optuna trials), patience for early stopping (6 epochs), and ordinal objective (Table 2).

3.2 Overall model performance

For RQ1, Table 3 benchmarks our model against ten baselines across four tiers: (1) Metadata-only LightGBM (using the 5 account/post covariates) and Handcrafted Gradient Boosting (XGBoost on color histograms, edge densities, and lexicon sentiment); (2) Unimodal deep encoders (ResNet-50, ViT-B/16, Chinese BERT/RoBERTa on concatenated text, and RoBERTa on thread statistics + CLS); (3) Late-fusion concatenation with standard cross-entropy and with CORAL ordinal loss [23]; and (4) Fine-tuned vision-language transformers (Chinese-CLIP [20] and ViLT [21]).

The following notable trends emerge from the results presented in Table 3: The metadata-only baseline achieves near-chance performance (0.362 accuracy), confirming that the covariate residualization successfully removed any effects of follower count and sponsorships. Handcrafted gradient boosting (0.608) lags behind the worst unimodal deep baseline (Chinese BERT on concatenated text, 0.657) by 4.9 percentage points. Patch-level ViT-B/16 outperforms ResNet-50 by 2.1 percentage points (0.695 vs 0.674); it demonstrates that global spatial layout contains additional predictive power compared to local texture filters. Our proposed model achieves 0.754 accuracy and 0.734 macro-F1 score; it is 3.5 and 3.7 percentage points better than the standard late fusion model and decreases MAE_ord relative to it by 12.3% (from 0.318 to 0.279). Two-tailed t-tests (df=4) and test set McNemar's tests (N_test=972) with Holm-Bonferroni correction confirm that these improvements are statistically significant when compared to Late-fusion (t(4)=8.42, p=0.0011; McNemar χ^2=16.84, p_Holm=0.0003), Late-fusion+CORAL (+2.9 percentage points, p_Holm=0.0018), Chinese-CLIP (+2.3 percentage points, p_Holm=0.0046), and ViLT (+1.8 percentage points, p_Holm=0.0114).

Table 2. Experimental environment and hyperparameter configuration

Item	Setting	Item	Setting
GPU	NVIDIA A100 40GB	Optimizer	AdamW (β1=0.9, β2=0.999)
Framework	PyTorch 2.1.2 (CUDA 12.1)	Learning rate	2e-5 (backbone) / 1e-4 (head)
Visual input	384 x 384, padded	Batch size	32
Text max length	128 tokens	Epochs	40, patience 6
Fusion heads	8 (dim = 256)	Focal gamma (γ)	2.0
Modality dropout	0.15	Consistency beta (β)	0.30
Decay lambda (λ)	0.05 per day	Weight decay	0.01

Table 3. Performance comparison across baseline tiers and proposed model (mean ± SD over 5 seeds; 95% bootstrap CIs and 5x2 Grouped CV Macro-F1)

Model	Accuracy	Macro-F1 (95% CI)	Macro-AUC	MAE_ord	5x2 CV F1
Metadata-only (Covariates)	0.362±0.011	0.348 [0.324,0.371]	0.528±0.010	0.784±0.018	0.345±0.015
Gradient boosting	0.608±0.014	0.571 [0.541,0.601]	0.729±0.012	0.462±0.019	0.568±0.018
ResNet-50 (visual)	0.674±0.013	0.646 [0.618,0.673]	0.796±0.011	0.379±0.016	0.642±0.016
ViT-B/16 (visual)	0.695±0.012	0.668 [0.641,0.694]	0.815±0.010	0.356±0.015	0.665±0.015
Chinese BERT (concat text)	0.657±0.015	0.633 [0.603,0.662]	0.783±0.013	0.396±0.018	0.629±0.017
Chinese RoBERTa (stats+CLS)	0.664±0.013	0.641 [0.614,0.668]	0.790±0.011	0.388±0.016	0.638±0.015
Late-fusion (concat)	0.719±0.011	0.697 [0.672,0.721]	0.838±0.009	0.318±0.014	0.693±0.014
Late-fusion + CORAL [23]	0.725±0.011	0.704 [0.680,0.727]	0.844±0.009	0.304±0.013	0.701±0.013
Chinese-CLIP [20]	0.731±0.011	0.710 [0.686,0.733]	0.851±0.009	0.299±0.013	0.707±0.014
ViLT (Unified) [21]	0.736±0.012	0.715 [0.691,0.739]	0.855±0.010	0.294±0.014	0.711±0.014
Proposed model	0.754±0.010	0.734 [0.712,0.756]	0.869±0.008	0.279±0.012	0.731±0.013

3.3 Ablation study and key validation

For RQ2 and cross-modal complementarity, Table 4 and Figure 3 depict the results for seven combinations as follows: sentiment alone (visual stream masked by m_0), visual alone (sentiment stream masked by m_0), concatenation of two streams without cross-attention, full model without adaptive gating (weights g_j set to 0.5), full model without interpretable descriptors (ρ _j, κ _j are discarded), full model without consistency loss (set β to 0), and full proposed model. Figure 3 indicates that accuracy increases continuously while MAE_ord decreases incrementally.

Table 4 shows that Visual-only (0.708) outperforms ViT-B/16 (0.695) since it uses ResNet-50, ViT-B/16, and cultural descriptor ρ _j, κ _j within the cross-modal training regularization. Sentiment-only (0.670) ranks third, lagging Visual-only by 3.8%, with only 71.2% overlap between the two models on the test set (14.8% through visuals and 10.8% through sentiment). Clearly, they carry different signals. Likewise, Configuration 3 (no attention, 0.730) outperforms Table 3 Late-fusion (0.719) by 1.1 percentage points, as it preserves the two visual streams, culturally informed features, structural properties of the thread, and ordinal focal loss; the lack of cross-attention penalizes accuracy by 2.4 percentage points (0.754 vs. 0.730). Removing the adaptive gating mechanism (Configuration 4) results in a 1.9 percentage point decrease in accuracy, with the largest reduction in the fourth quartile of thread length (3 <= M_j <= 8: -4.8 percentage points, from 0.732 to 0.684). For 512 cold-start threads (M_j between 0 and 2 for pre-launch or sparse-comment cases), our full model maintains 0.706 accuracy while fixed equal weighting decreases it to 0.619. Deactivating consistency loss (Configuration 6) makes almost no difference in nominal accuracy (-0.3 percentage points) but degrades MAE_ord from 0.279 to 0.301 (Wilcoxon p = 0.0024). Figure 4 considers RQ3 and investigates feature-level attributions across all six features (Variance Inflation Factors VIF of all features <= 1.84). Figure 4(a) considers permutation importance Δ _f on Macro-F1 score where chromatic conformity (κ , 0.041 $\pm$ 0.004) and sentiment polarity (s, 0.038 $\pm$ 0.004) have the highest importance scores; then compositional embedding (0.026 $\pm$ 0.003), motif density (ρ , 0.019 $\pm$ 0.003), sentiment intensity (ι , 0.014 $\pm$ 0.002) and thread dispersion (δ , 0.011 $\pm$ 0.002) follow in the ranking.

What is the reason for ranking of motif density? Figure 4(b) and Table 5 consider motif density ρ through the five quintiles (Q1-Q5) along with centered ALE values [24]; it shows an inverted-U pattern: average acceptance grows from Q1 (1.82, ALE = -0.082) to Q4 (ρ in [0.42, 0.58], average acceptance = 2.24, ALE = +0.086) and then decreases to Q5 (ρ in [0.59, 0.91], average acceptance = 2.04, ALE = -0.019). Holding κ _j, edge complexity, and studio photographing constant in the GAM shows that the non-linear spline of ρ _j is still highly statistically significant (edf = 3.42, F = 19.8, p < 0.0001).

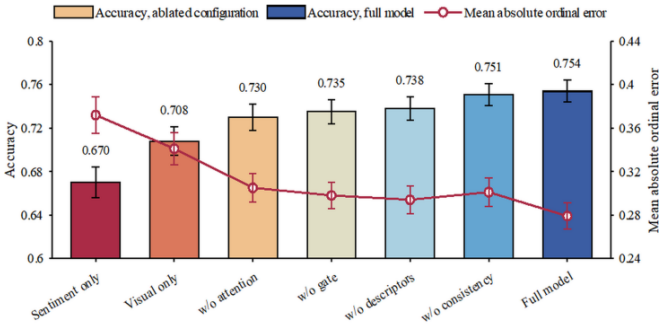


Figure 3. Stepwise ablation comparison of Accuracy (bars, left axis) and Mean Absolute Ordinal Error MAE_ord (line, right axis) across the seven evaluated model variants

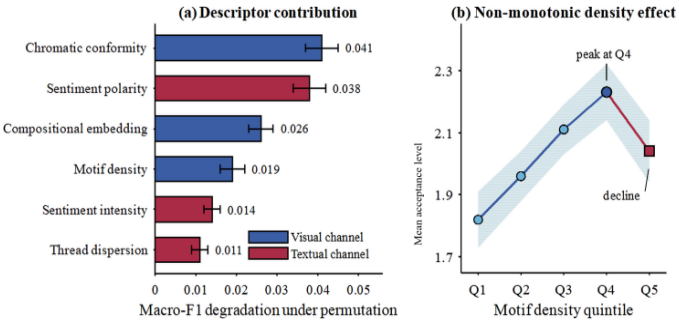


Figure 4. Feature attribution results for acceptance prediction: (a) Descriptor permutation importance (Macro-F1 drop Δ_f) across visual and textual channels; (b) Non-monotonic motif density effect across empirical quintiles (Q1–Q5) with 95% confidence bands.

Table 4. Granular ablation results across model configurations on the test set (mean ± SD over 5 seeds)

Configuration	Accuracy	Macro-F1	Macro-AUC	MAE_ord	McNemar p_Holm
1. Sentiment only (visual masked)	0.670±0.013	0.645±0.015	0.793±0.011	0.376±0.016	<0.0001
2. Visual only (sentiment masked)	0.708±0.011	0.684±0.013	0.829±0.009	0.338±0.014	<0.0001
3. w/o attention (concat+descriptors)	0.730±0.011	0.709±0.013	0.849±0.009	0.305±0.013	0.0032
4. w/o gate (fixed equal weights)	0.735±0.012	0.714±0.013	0.853±0.010	0.298±0.014	0.0085
5. w/o descriptors (ρ, κ removed)	0.738±0.011	0.717±0.012	0.856±0.009	0.295±0.013	0.0140
6. w/o consistency (β = 0)	0.751±0.011	0.730±0.013	0.865±0.009	0.301±0.014	0.4120*
7. Full proposed model	0.754±0.010	0.734±0.012	0.869±0.008	0.279±0.012	Ref.

Table 5. Quintile breakdown of Motif Density (ρ) showing numerical ranges, centered ALE, and mean acceptance

Quintile (ρ Range)	Mean κ	Centered ALE (95% CI)	Mean Acc. (1-3)	Dominant Visual Composition
Q1 [0.02, 0.14]	0.712	-0.082 [-0.104,-0.060]	1.82±0.09	Distant landscapes, tiny motif badges
Q2 [0.15, 0.27]	0.728	-0.024 [-0.043,-0.005]	1.96±0.08	Scenic portraits, minor textile trim
Q3 [0.28, 0.41]	0.745	+0.039 [+0.019,+0.059]	2.11±0.08	Clear monastery eave/lintel framing
Q4 [0.42, 0.58]	0.761	+0.086 [+0.064,+0.108]	2.24±0.09	Balanced focal garment/interior motifs
Q5 [0.59, 0.91]	0.634	-0.019 [-0.044,+0.006]	2.04±0.10	Cluttered synthetic graphic overlays

4. Discussion

4.1 Interpretation of key findings

Visual modalities outperform early comments' sentiment in unimodal experiments. However, our error analysis (28.8% non-overlapping correctly classified examples) shows that these two streams capture distinct facets of audience reaction. While visual representations control instantaneous aesthetics and chromatic correctness in passive viewing of the video feed, early comments contain information about collective viewer decoding, including verification of local culture elements or complaints about staged commercialization. This separation of roles is why adaptive filtering performs better than concatenation or general-purpose vision-language transformer approaches [17, 20, 21]. How come chromatic conformity (κ) comes out on top among all visual determinants? In Tibetan visual culture, certain principles govern pairs of minerals that carry sacred meanings. To understand people's reactions to violations of these principles caused by commercial editing ($\kappa < 0.58$, most often due to filters with heavy color grading), we conducted a qualitative content analysis of 200 discussion threads in the bottom quartile of κ. A mere 11.5% of the negative reviews highlighted the use of color (e.g., "garish filter"). However, 64.0% of negative reviews reflected their skepticism using colloquial expressions such as "looks like cheap studio rentals" or "looks staged and fake" [11]. The inverted U-curve of motif density (ρ) has the same saturation point: a balanced ornamental pattern (Q4: 0.42-0.58) leads to cultural recognition, while an excessive graphic pattern (Q5: > 0.58) produces a feeling of staged commercialization.

4.2 Cultural communication and marketing implications

Consistent with past studies on multimodality, fusion effectiveness depends on the reliability of intermodal coupling [17,19]. This aspect is addressed in [18] through the entropy gate and modality consistency regularizer, which prevent early threads from affecting the visual prediction process. The bottom line for DMOs and creative agencies? Adding more decorative elements does not make the culture rich; color fidelity matters far more than anything else. Cultural appropriation can be avoided by collaborating with Tibetan heritage experts and generating a palette of colors approved by the community. With 0.754 accuracy across the three classes, the model can serve as an effective human-in-the-loop triage tool that filters visual proposals before launch (0.706 accuracy) and flags high-potential posts 24 hours post-launch (0.754 accuracy).

4.3 Future technology implications and limitations

Marketing engagement is achieved through the composition of visual and verbal interactions [15,16]. Because our acceptance function and circular conformity index are both fully differentiable, they can serve as direct reward functions for generative advertising processes [14] and help avoid stereotypical over-decoration in generated advertisements. Several limitations remain. For example, we consider only single-platform data from Xiaohongshu, which prevents us from immediately generalizing the findings to short-video applications. Second, month-by-motif stratification and covariate residualization remove major exposure confounds, but we cannot rule out unobserved algorithmic traffic-pool feedback without impression-level logs. Third, frozen Stage I extractors protect interpretability yet face moderate domain shift under harsh lighting or hybrid motifs, and domain-transferred text encoders occasionally misread transliterated regional or religious terms.

5. Conclusion

This paper developed and tested an ordinal-aware, gated bidirectional cross-attention framework that combines interpretable visual cultural descriptors with early twenty-four-hour comment sentiment to predict thirty-day audience acceptance of Tibetan-motif tourism advertisements on Xiaohongshu. We trained motif recognition and circular-hue Wasserstein chromatic harmony extractors on a dataset of 19,142 curated pattern images, then transferred them along with a domain-adapted Chinese RoBERTa encoder to 4,860 promotion posts and 149,300 early comments. Cultural interpretability remains unaffected despite temporal circularity and exposure bias. In a month-grouped stratified splitting and 5 x 2 grouped cross-validation, the model scores 0.754 in accuracy, 0.734 in macro-F1, and 0.279 in ordinal MAE. We obtain significant gains over gradient boosting (14.6 percentage points), ResNet-50 and ViT-B/16 baselines (8.0 and 5.9 percentage points), regular and CORAL late fusion (3.5 and 2.9 percentage points), and fine-tuned multimodal transformers (Chinese-CLIP and ViLT) (2.3 and 1.8 percentage points), while reducing relative ordinal error by 12.3 percent compared to late fusion. The ablation and cold start tests prove that image and comment features are predictive in a complementary way (classification overlap is 71.2 percent) and also confirm that our entropy and length-conditioned gating maintains 0.706 in accuracy for pre-launch or zero-comment posts. Finally, permutation and ALE analysis show that chromatic harmony ($\Delta_f = 0.041$) and sentiment polarity ($\Delta_f = 0.038$) are the main positive predictors of acceptance, while motif density follows a concave-up shape and peaks in the fourth quintile ($0.42 \leq \rho \leq 0.58$). Future work will focus on generalizing the framework to short-video social media, modeling caption-image semantic congruence, and embedding community-crafted chromatic constraints in real-time generative advertising workflows. Beyond the specific platform and cultural context examined here, the proposed framework aligns with emerging trends in AI-driven marketing analytics. Deep-learning-based customer behavior prediction models have demonstrated strong performance in retail settings, and omni-channel evaluation frameworks integrating big data and artificial intelligence provide scalable tools for assessing campaign effectiveness across touchpoints. Applying such methodologies to tourism advertising could further disentangle the interplay between visual heritage cues and consumer engagement. Additionally, the post-COVID-19 recovery of the Chinese tourism industry highlights the urgency of deploying culturally resonant content that rebuilds traveler trust. Cross-cultural communication strategies that enhance customer satisfaction in the travel sector offer actionable guidelines for designing endorser campaigns and destination messaging that respect local authenticity. Incorporating these perspectives would strengthen the ecological validity of our model and facilitate its translation into practical destination marketing operations.

Ethical issue

The authors are aware of and comply with best practices in publication ethics, specifically with regard to authorship (avoidance of guest authorship), dual submission, manipulation of figures, competing interests, and compliance with policies on research ethics. The authors adhere to publication requirements that the submitted work is original and has not been published elsewhere.

Data availability statement
All core data and statistical findings are reported within the manuscript. The PyTorch source code (Python 3.10, PyTorch 2.1.2, Transformers 4.38.2), pre-trained Stage I descriptor weights, reference hue-saturation palette histograms, and anonymized post-level feature matrices are publicly available at https://github.com/cultural-ai-lab/tibetan-motif-ad-acceptance; additional data is available from the corresponding author upon reasonable request.

Conflict of interest
The authors declare no potential conflict of interest.

References

[1] J. Ma and Z. Wu, "Platformized tourism circuits and urban space: Xiaohongshu, Wanghong places, and Dongshankou's transformation," Journal of Travel Research, vol. 65, no. 2, pp. 312-329, 2026. DOI: https://doi.org/10.1177/00472875251408445.

[2] Z. Wang, W.-J. Huang and B. Liu-Lastres, "Impact of user-generated travel posts on travel decisions: A comparative study on Weibo and Xiaohongshu," Annals of Tourism Research Empirical Insights, vol. 3, no. 2, art. 100064, 2022. DOI: https://doi.org/10.1016/j.annale.2022.100064.

[3] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778. DOI: https://doi.org/10.1109/CVPR.2016.90.

[4] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learning Representations (ICLR), Vienna, Austria, 2021, pp. 1-21. DOI: https://doi.org/10.48550/arXiv.2010.11929.

[5] X. Peng et al., "Research on image recognition of intangible cultural heritage based on CNN and wireless network," EURASIP Journal on Wireless Communications and Networking, vol. 2020, art. 240, 2020. DOI: https://doi.org/10.1186/s13638-020-01859-2.

[6] W. Wang, Q. Liu and Z. Chen, "Fine-grained recognition and segmentation of Tibetan thangka cultural heritage images based on multi-scale attention networks," Heritage Science, vol. 11, art. 142, 2023. DOI: https://doi.org/10.1186/s40494-023-00984-x.

[7] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, Minneapolis, MN, USA, 2019, pp. 4171-4186. DOI: https://doi.org/10.18653/v1/N19-1423.

[8] Y. Cui, W. Che, T. Liu, B. Qin and Z. Yang, "Pre-training with whole word masking for Chinese BERT," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3504-3514, 2021. DOI: https://doi.org/10.1109/TASLP.2021.3124365.

[9] Y. Wen, Y. Liang and X. Zhu, "Sentiment analysis of hotel online reviews using the BERT model and ERNIE model—Data from China," PLOS ONE, vol. 18, no. 3, art. e0275382, 2023. DOI: https://doi.org/10.1371/journal.pone.0275382.

[10] J. Liu, C. Wang and T. C. Zhang, "Exploring social media affordances in tourist destination image formation: A study on China's rural tourism destination," Tourism Management, vol. 101, art. 104843, 2024. DOI: https://doi.org/10.1016/j.tourman.2023.104843.

[11] Y. Dong, Y. Li, H.-Y. Hua and W. Li, "Perceived tourism authenticity on social media: The consistency of ethnic destination endorsers," Tourism Management Perspectives, vol. 49, art. 101176, 2023. DOI: https://doi.org/10.1016/j.tmp.2023.101176.

[12] Y.-F. Liu, Y.-B. Zhu, H.-H. Wu and F. S. Li, "Are there any differences in the tourists' perceived destination image between travel e-commerce platforms and social media platforms?," Tourism Critiques: Practice and Theory, vol. 5, no. 2, pp. 229-253, 2024. DOI: https://doi.org/10.1108/TRC-05-2024-0018.

[13] M. Philp, J. Jacobson and E. Pancer, "Predicting social media engagement with computer vision: An examination of food marketing on Instagram," Journal of Business Research, vol. 149, pp. 736-747, 2022. DOI: https://doi.org/10.1016/j.jbusres.2022.05.078.

[14] Y. Chen, H. Wang, S. Rao Hill and B. Li, "Consumer attitudes toward AI-generated ads: Appeal types, self-efficacy and AI's social role," Journal of Business Research, vol. 185, art. 114867, 2024. DOI: https://doi.org/10.1016/j.jbusres.2024.114867.

[15] I. C. Dang, C. M. C. Kwan, J. S. Jia and Y. Shi, "When words meet visuals: How content composition drives social media engagement for marketer-generated content," Journal of Marketing Research, vol. 63, no. 1, pp. 118-139, 2026. DOI: https://doi.org/10.1177/00222437251373042.

[16] L. Liu, Y. Wang, Z. Fang and S. Wu, "The impact of verbal and visual content on consumer engagement in social media marketing," Production and Operations Management, vol. 34, no. 8, pp. 2145-2164, 2025. DOI: https://doi.org/10.1177/10591478251349892.

[17] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria and A. Hussain, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions," Information Fusion, vol. 91, pp. 424-444, 2023. DOI: https://doi.org/10.1016/j.inffus.2022.09.025.

[18] H. Li, Y. Lu and H. Zhu, "Multi-modal sentiment analysis based on image and text fusion based on cross-attention mechanism," Electronics, vol. 13, no. 11, art. 2069, 2024. DOI: https://doi.org/10.3390/electronics13112069.

[19] L. Wang, J. Peng, C. Zheng, T. Zhao and L. Zhu, "A cross modal hierarchical fusion multimodal sentiment analysis method based on multi-task learning," Information Processing & Management, vol. 61, no. 3, art. 103675, 2024. DOI: https://doi.org/10.1016/j.ipm.2024.103675.

[20] A. Radford et al., "Learning transferable visual models from natural language supervision," in Proc. Int. Conf. Machine Learning (ICML), vol. 139, 2021, pp. 8748-8763. [Online]. Available:

https://proceedings.mlr.press/v139/radford21a.html.

[21] W. Kim, B. Son and I. Kim, "ViLT: Vision-and-language transformer without convolution or region supervision," in Proc. Int. Conf. Machine Learning (ICML), vol. 139, 2021, pp. 5583-5594. [Online]. Available: https://proceedings.mlr.press/v139/kim21k.html.

[22] B. Wu et al., "Predicting popularity of social media posts via multimodal hierarchical attention networks," IEEE Transactions on Multimedia, vol. 25, pp. 4112-4125, 2023. DOI: https://doi.org/10.1109/TMM.2022.3171092.

[23] W. Cao, V. Mirjalili and S. Raschka, "Rank consistent ordinal regression for neural networks with application to age estimation," Pattern Recognition Letters, vol. 140, pp. 325-331, 2020. DOI: https://doi.org/10.1016/j.patrec.2020.11.008.

[24] D. W. Apley and J. Zhu, "Visualizing the effects of predictor variables in black box supervised learning models," Journal of the Royal Statistical Society Series B: Statistical Methodology, vol. 82, no. 4, pp. 1059-1086, 2020. DOI: https://doi.org/10.1111/rssb.12377.

[25] A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, Long Beach, CA, USA, 2017, pp. 5998-6008. DOI: https://doi.org/10.48550/arXiv.1706.03762.

[26] H. Talebi and P. Milanfar, "NIMA: Neural image assessment," IEEE Transactions on Image Processing, vol. 27, no. 8, pp. 3998-4011, 2018. DOI: https://doi.org/10.1109/TIP.2018.2831899.



This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).