

Article

A Multimodal Analysis of Automotive Video Communication Effectiveness: The Impact of Visual Emotion, Spatiotemporal Cues, and Title Sentiment

Yawei He ^{1,*}, Zijie Feng ^{1,2,†} and Wen Liu ³ ¹ School of Media and Communication, Shanghai Jiao Tong University, Shanghai 200240, China² School of Computer Technology and Application, Qinghai University, Xining 810016, China³ AI Technology Innovation Group, School of Economics and Management, Communication University of China, Beijing 100024, China

* Correspondence: heyawei1992@sjtu.edu.cn

† These authors contributed equally to this work.

Abstract

To quantify the communication effectiveness of automotive online videos, this study constructs a multimodal deep learning framework. Existing research often overlooks the intrinsic and interactive impact of textual and dynamic visual content. To bridge this gap, our framework conducts an integrated analysis of both the textual (titles) and visual (frames) dimensions of videos. For visual analysis, we introduce FER-MA-YOLO, a novel facial expression recognition model tailored to the demands of computational communication research. Enhanced with a Dense Growth Feature Fusion (DGF) module and a multiscale Dilated Attention Module (MDAM), it enables accurate quantification of on-screen emotional dynamics, which is essential for testing our hypotheses on user engagement. For textual analysis, we employ a BERT model to quantify the sentiment intensity of video titles. Applying this framework to 968 videos from the Bilibili platform, our regression analysis—which modeled four distinct engagement dimensions (reach, support, discussion, and interaction) separately, in addition to a composite effectiveness score—reveals several key insights: emotionally charged titles significantly boost user interaction; visually, the on-screen proportion of human elements positively predicts engagement, while excessively high visual information entropy weakens it. Furthermore, neutral expressions boost view counts, and happy expressions drive interaction. This study offers a multimodal computational framework that integrates textual and visual analysis and provides empirical, data-driven insights for optimizing automotive video content strategies, contributing to the growing application of computational methods in communication research.

Keywords: multimodal analysis; computational communication; facial expression recognition; communication effectiveness; video content analysis; emotional intensity; user engagement



Academic Editor: Manohar Das

Received: 30 September 2025

Revised: 14 October 2025

Accepted: 22 October 2025

Published: 27 October 2025

Citation: He, Y.; Feng, Z.; Liu, W. A Multimodal Analysis of Automotive Video Communication Effectiveness: The Impact of Visual Emotion, Spatiotemporal Cues, and Title Sentiment. *Electronics* **2025**, *14*, 4200. <https://doi.org/10.3390/electronics14214200>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. 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/>).

1. Introduction

In the era of rapid development of digital media, online video platforms have emerged as the core battleground for information dissemination and brand promotion, triggering profound changes in the content ecosystem [1]. In this wave, the automotive industry is accelerating its transition from traditional advertising formats to video-based marketing

strategies [2]. Automotive-related online videos, with their vivid and intuitive visual expression and emotional interaction characteristics, have become an important medium for brands to connect with users and influence consumer decisions.

Despite the explosive growth of automotive videos on mainstream platforms, there remains a significant lack of understanding regarding their dissemination mechanisms. While some foundational work has begun to explore the multimodal elements within specific automotive advertisements (e.g., Izadifar, P. conducted qualitative analysis on a Volkswagen car advertisement video [3]), current analyses are often limited to superficial factors such as titles, tags, or posting times, or rely on qualitative assessments based on subjective experience [4]. Crucially, existing multimodal studies often struggle to obtain quantitative insights into viewer perception and engagement. This research vacuum directly leads to strategic blindness and resource misallocation in marketing practices, making it difficult for brands to effectively optimize video content to enhance communication performance. This exposes a critical gap: in the dynamic narrative of videos, how do the core visual components—characters as emotional and narrative subjects, and vehicles as product focal points—interact in terms of their spatiotemporal arrangement and emotional expression to influence dissemination performance (e.g., views, likes, and shares)? For instance, Emotional Contagion Theory posits that characters' positive emotions can elicit audience resonance, while Cognitive Load Theory suggests that the spatiotemporal arrangement of visual elements directly impacts the efficiency of information reception. Crucially, this limitation not only stems from methodological gaps, but also from the lack of vision models capable of reliably measuring theoretically grounded visual constructs—such as the temporal ratio of natural expressions or human–vehicle spatial co-occurrence—at scale. This hinders the transition of marketing strategies toward intelligent, data-driven approaches [5].

To bridge this research gap, this paper constructs a deep learning-supported multimodal video analysis system, decomposing videos into temporally and spatially measurable element sequences through a comprehensive framework that systematically examines both visual and textual cues for video dissemination effectiveness. Specifically, we employ advanced object detection algorithms to precisely locate the appearance intervals and interaction patterns of characters and vehicles. Our analytical approach follows a two-stage regression methodology: first, we conduct regression analysis using visual cues extracted from video content, including human–vehicle spatiotemporal patterns and emotional expressions; subsequently, we perform additional regression analysis incorporating textual sentiment features derived from video titles to provide a comprehensive understanding of communication effectiveness. Based on approximately 1000 automotive video samples obtained from the Bilibili platform, this study conducts empirical validation to systematically describe the spatio-temporal distribution patterns of characters and vehicles in automotive videos, particularly their co-occurrence duration and proportional relationships within the overall structure. After controlling for confounding factors such as video duration and posting timepoints, we utilize multivariate regression and machine learning tools [6] to deeply examine the intrinsic connections between human–vehicle interaction patterns, emotional intensity, variability, and dissemination metrics, while simultaneously analyzing the sentiment polarity and emotional intensity of video titles as complementary predictive factors. We develop and validate a reliable predictive framework to uncover the visual patterns driving high-dissemination videos and provide empirical guidance for marketing optimization [7].

To our knowledge, this study is among the first to apply deep visual analytics technology in the context of automotive digital marketing, enhancing analytical precision and depth through quantitative decomposition of the spatiotemporal patterns of key video

elements [8]. Second, this paper provides data-driven insights for content creators and marketing practitioners, shifting decision-making from experience-based reliance to precise empirical evidence derived from both visual content analysis and textual feature extraction. Through our proprietary optimization model FER-MA-YOLO (details shown in Figure 1), this research contributes to extending the application of visual communication theory in industry and provides preliminary theoretical and experimental support for AI-assisted content effectiveness evaluation [9]. The contributions of this paper can be summarized as follows:

- This study introduces a deep learning-driven multimodal video analysis framework to the field of automotive digital marketing research. The framework is designed to quantitatively decode the spatiotemporal interaction patterns between people and vehicles in online videos and to explore their nonlinear impact on dissemination metrics such as views, likes, and shares.
- We propose a network architecture utilizing dense connections to enhance the efficiency of multiscale feature fusion, which we term the Dense Growth Feature Fusion (DGF) module. This module is designed to improve accuracy in tasks such as person-vehicle co-occurrence detection and facial expression quantification, while maintaining real-time efficiency.
- We propose a dual attention mechanism, termed the Multiscale Dilated Attention Module (MDAM), which is based on multiscale dilated convolutions. Its purpose is to achieve adaptive feature weighting, thereby enhancing the model's adaptability and recognition capabilities for features of various scales in images.
- We designed the FER-MA-YOLO model and conducted comprehensive empirical comparisons with several advanced object detection methods. On the Facial Expression Recognition-9 Emotions dataset, our model's detection performance outperformed the compared baseline methods, as shown in Figure 2.

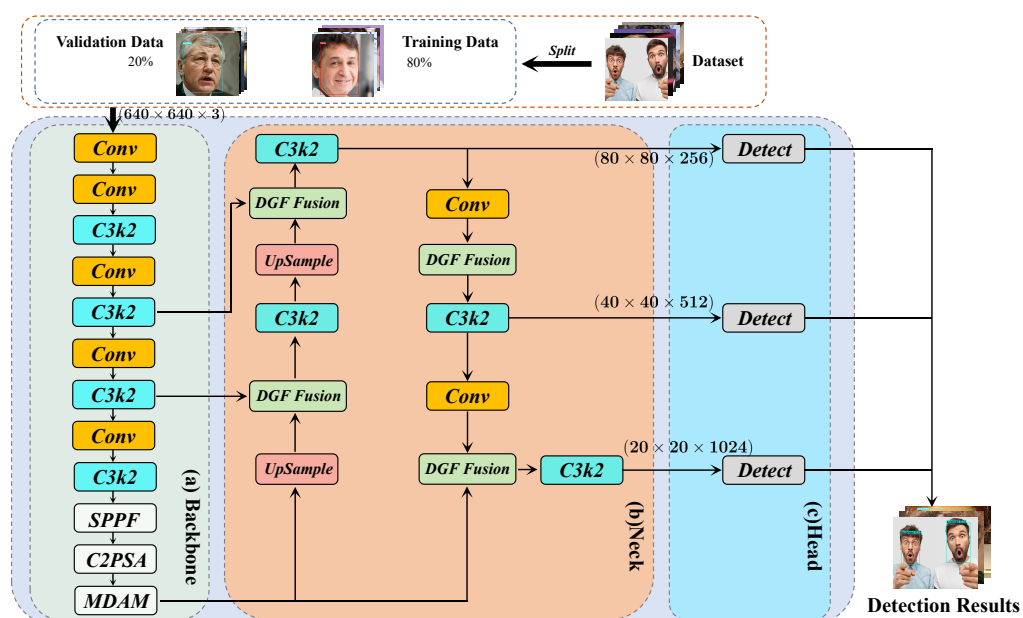


Figure 1. The architecture of the proposed FER-MA-YOLO.

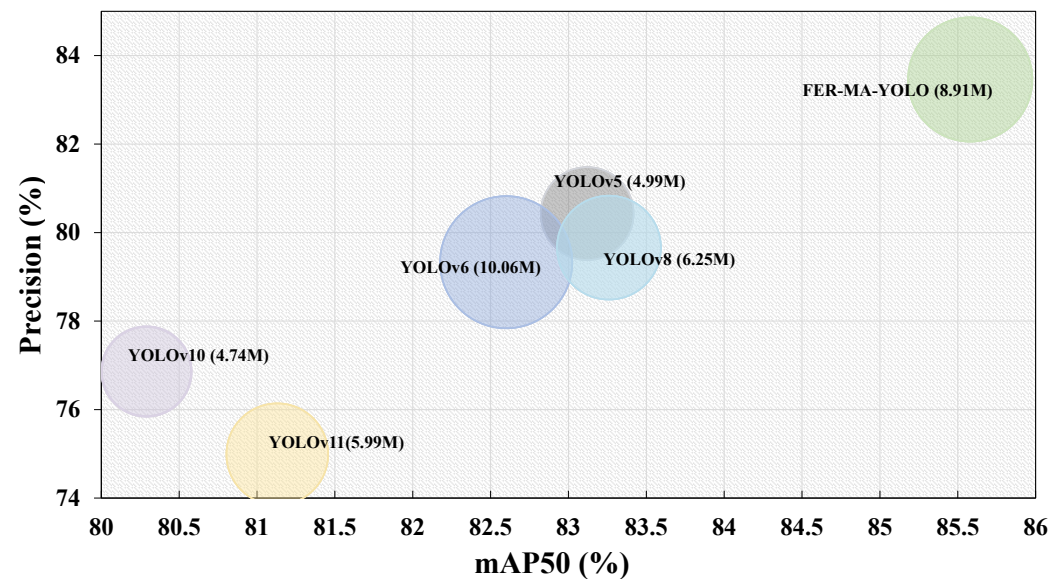


Figure 2. Comparative analysis of the proposed FER-MA-YOLO model against baseline YOLO models on the facial expression recognition dataset. The horizontal axis represents mAP50 (%), while the vertical axis displays Precision (%). The size of each data point corresponds to the number of model parameters (in millions, M), which is explicitly labeled next to each bubble to ensure precise interpretation of the performance–complexity trade-off.

2. Research Variables and Hypotheses

This project is based on visual communication theory and a multimodal content analysis framework, using car-related videos on the Bilibili platform as the research object. While recent work has applied qualitative multimodal analysis to individual automotive advertisements, our study scales this inquiry through deep learning to quantitatively extract key feature variables reflecting the core content of videos across a large dataset. Using the YOLO object detection algorithm and facial expression recognition technology, a multi-dimensional feature system was constructed that includes human emotional expression, human–vehicle spatiotemporal distribution, and visual proportion.

(1) Characteristics of emotional expression in characters: Emotions are an important factor influencing the audience’s psychological responses and behavioral intentions. In automotive videos, the emotional state of the host or on-screen characters directly impacts viewers’ viewing experiences and willingness to engage. This study identifies three dimensions—the proportion of time spent in a happy emotional state, the proportion of time spent in a natural state, and the proportion of time spent in a surprised emotional state—to explore the differentiated effects of different emotional types on communication outcomes. Therefore, Hypotheses H1–H3 are proposed.

(2) Spatial and temporal distribution characteristics of characters: As the primary communication subjects in videos, the presentation of characters in temporal and spatial dimensions significantly influences the audience’s attention allocation. The proportion of time that characters appear reflects their participation in the temporal dimension, while the proportion of screen area occupied by characters reflects their importance in the visual space. These two dimensions collectively constitute the spatial and temporal presentation patterns of character elements. Therefore, Hypotheses H4 and H5 are proposed.

(3) Vehicle presentation characteristics: In automotive videos, vehicles serve as the core display objects, and their presentation directly impacts the professionalism and appeal of the content. The proportion of vehicle appearance duration and the proportion of vehicle screen area respectively measure the importance of vehicle elements from the perspec-

tives of temporal continuity and visual prominence. Therefore, Hypotheses H6 and H7 are proposed.

(4) Comprehensive emotional intensity index of titles: In addition to specific emotional types, the overall intensity of emotional expression in titles may also influence dissemination effects. Emotional intensity scores comprehensively reflect the vividness and persuasiveness of character emotional expressions. Therefore, we propose hypothesis H8.

(5) The scope of human visual perception: We aim to explore whether visual differences in brightness and image entropy can attract audience attention, influence visual perception memory, and thereby impact video dissemination effectiveness. Specifically, differences in brightness may capture audience attention, while higher image entropy values typically indicate richer information and details, which may enhance video dissemination effectiveness. Based on this research approach, we propose Hypotheses H9 and H10.

Based on the above analysis, this study proposes the following hypotheses:

Hypothesis 1 (H1). *In Bilibili automotive-related videos, the proportion of happy emotions has a positive impact on video dissemination effectiveness.*

Hypothesis 2 (H2). *In Bilibili automotive-related videos, the proportion of natural emotions has a positive impact on video dissemination effectiveness.*

Hypothesis 3 (H3). *In Bilibili automotive-related videos, the proportion of surprise-related emotional content duration has a negative impact on video dissemination effectiveness, as this emotion may introduce uncertainty or be associated with unexpected negative events, potentially reducing viewer confidence in a high-involvement product category.*

Hypothesis 4 (H4). *In Bilibili automotive-related videos, the proportion of character appearance duration has a positive impact on video dissemination effectiveness.*

Hypothesis 5 (H5). *In Bilibili automotive-related videos, the proportion of character screen area has a positive impact on video dissemination effectiveness.*

Hypothesis 6 (H6). *In Bilibili automotive-related videos, the proportion of vehicle appearance duration has a positive impact on video dissemination effectiveness.*

Hypothesis 7 (H7). *In Bilibili automotive-related videos, the proportion of vehicle screen area has a positive impact on video dissemination effectiveness.*

Hypothesis 8 (H8). *In Bilibili automotive-related videos, the emotional intensity of the title has a positive impact on video dissemination effectiveness.*

Hypothesis 9 (H9). *In Bilibili automotive-related videos, the video's information entropy has a negative impact on video dissemination effectiveness, as excessively high visual complexity may increase cognitive load and hinder information processing.*

Hypothesis 10 (H10). *In Bilibili automotive-related videos, the average brightness of the video has a positive impact on video dissemination effectiveness.*

In addition to the above core feature variables, this study also collected video metadata as control variables, including video length, publication time, title features, etc., to ensure the reliability and validity of the research results. In order to explore the impact of different emotional expressions and their duration ratios on audience behavior in automotive video marketing, we propose the following research questions:

- Q1: In automotive video marketing, what is the optimal duration ratio for expressing human emotions? How can emotional ratio control be used to increase audience engagement?

3. Research Design and Data Operationalization

3.1. Automotive Video Data Sources

This study primarily utilizes automotive-themed videos from the Bilibili platform as its core sample source, encompassing content such as vehicle reviews, test drive experiences, model comparisons, and brand events. Video data was collected from July 2019 to December 2024, yielding an initial sample size of approximately 1050 videos. During data cleaning, samples that could potentially impact feature extraction and analysis were excluded. These included videos with subpar image quality, missing information about people or vehicles, and audio–visual desynchronization. Ultimately, $N = 968$ videos were retained as the research dataset. These videos have view counts ranging from tens of thousands to millions, representing creators of varying scales and audience levels. This diversity provides high representativeness for subsequent analysis.

3.2. Rationale for Platform Selection and Data Validation

The Bilibili platform was selected as the data source for three primary reasons. First, its unique “bullet comment” (danmaku) system fosters a distinctive, real-time interactive community, providing a rich and multi-dimensional measure of user engagement that is central to our definition of “communication effectiveness.” Second, Bilibili has become the dominant platform for in-depth, long-form automotive content in China, hosting a diverse and representative corpus of videos, ranging from professional reviewers to official brand channels. Third, studying communication dynamics on this platform offers significant insights into the world’s largest automotive market.

Regarding data validation, our dependent variables (e.g., view count, likes, comments) were directly collected from the Bilibili API, ensuring objective measurement. The independent variables were extracted via an automated pipeline using the FER-MA-YOLO and BERT models. The validity of this pipeline is supported by the state-of-the-art performance of FER-MA-YOLO on the benchmark FER-9 dataset (see Section 6) and the widespread acceptance of the BERT model for textual sentiment analysis in the NLP community.

3.3. Automotive Video Feature Extraction

Given that this study focuses on the spatiotemporal distribution characteristics of human emotional expressions and vehicle presentation methods—constructs central to our theoretical hypotheses—the video was first parsed frame-by-frame into static image frames. To ensure the validity of our operationalization, we employ FER-MA-YOLO, a purpose-built model whose architecture (e.g., DGF for stable temporal feature fusion, MDAM for subtle expression sensitivity) directly addresses the measurement challenges inherent in quantifying these communication variables from real-world automotive videos. In the 968 automotive marketing videos analyzed, “happy,” “neutral,” and “surprised” constitute the overwhelming majority of all facial expressions, while negative emotions (e.g., anger, sadness, disgust, fear) are virtually absent—likely due to normative constraints in brand communication. Therefore, focusing on these dominant and theoretically meaningful categories enhances both model robustness and interpretability, which are essential for applied communication research. Therefore, by integrating a facial expression recognition model (trained on the FER-9 dataset), we extract the duration proportions of three core emotions: “happy,” “neutral,” and “surprised”. At the vehicle feature level, the temporal proportion and visual area coverage of vehicles within the video were calculated. Additionally, brightness and image entropy were quantified to reflect the information density and detail

richness of the visual scene. All features underwent normalization and temporal alignment processing to ensure comparability of feature values across different videos.

3.4. User Visual Feature Quantification

Information Entropy: This was used to measure the uncertainty or randomness of a data source. This study first employed the OpenCV library to convert all sampled frames into grayscale images, thereby reducing the dimensionality of image information and enhancing computational processing speed. High entropy values indicate complex, detail-rich images (e.g., textured scenes), while low entropy values suggest uniformity (e.g., blank or monotonous areas). In video processing, this metric is particularly effective for detecting transition effects like cuts or fades by analyzing entropy fluctuations between frames. For grayscale images, entropy H is calculated based on the probability distribution of pixel intensities. A frame with entirely black pixels yields an entropy value of 0, while highly textured frames exhibit higher entropy values (not exceeding a maximum of 8). This is because an 8-bit grayscale image possesses $L = 256$ intensity levels (ranging from 0 to 255). The calculation formula is as follows:

$$H = - \sum_{i=0}^{L-1} p_i \log_b(p_i)$$

where $p_i = \frac{h_i}{N}$ denotes the probability of intensity level i , h_i denotes the pixel frequency with intensity i , and N is the total number of pixels.

Image Luminance: Unlike grayscale intensity, which can conflate color with brightness, luminance in a color space separates brightness from chromaticity components, enabling more refined analysis. In video frames, luminance variations can indicate changes in lighting, exposure inconsistencies, or environmental factors—critical for lighting normalization and object detection under variable conditions. The Hue–Saturation–Value (HSV) color model provides an intuitive representation for brightness extraction. HSV transforms the additive RGB color space into a cylindrical coordinate system where the Value (V) explicitly encodes luminance. The conversion from normalized RGB values ($R, G, B \in [0, 1]$) to HSV is defined by the following Equation (a)–(c):

$$H = \begin{cases} 0^\circ & \text{if } \Delta = 0, \\ 60^\circ \times \left(\frac{G-B}{\Delta} \bmod 6 \right) & \text{if } M = R \\ 60^\circ \times \left(\frac{B-R}{\Delta} + 2 \right) & \text{if } M = G \\ 60^\circ \times \left(\frac{R-G}{\Delta} + 4 \right) & \text{if } M = B \end{cases} \quad (a)$$

$$S = \begin{cases} 0 & \text{if } M = 0 \\ \frac{\Delta}{M} & \text{else} \end{cases} \quad (b)$$

$$V = M \quad (c)$$

where $M = \max(R, G, B)$, $m = \min(R, G, B)$, and $\Delta = M - m$

Brightness is then calculated as the mean value across all pixels in the V channel: $\text{Brightness} = \frac{1}{N} \sum_{i=1}^N V_i$, where V_i is the value component of the i -th pixel (typically scaled to $[0, 255]$ in 8-bit images). Information entropy quantifies image texture complexity and information content based on grayscale distribution. Image brightness evaluates global illumination levels through the V component of HSV. Entropy focuses on structural diversity, while brightness emphasizes illumination variation, jointly supporting scene detection and quality assessment in video frame analysis.

3.5. Video Title Sentiment Intensity Quantification

To delve into the mechanism by which video titles influence user engagement, this study employs a deep learning-based sentiment intensity analysis method to quantitatively assess the emotional orientation of video titles. Considering the diversity of automotive

video content and the complexity of title expressions, a pre-trained BERT multilingual sentiment classification model serves as the core analytical tool. Trained on a five-star rating system, this model effectively captures shifts in sentiment polarity and intensity within text. In the implementation, AutoTokenizer first preprocesses title texts through word segmentation, padding, and truncation to standardize input length to 128 tokens for model compatibility. Subsequently, the pre-trained BERT model calculates probability distributions across five sentiment levels for each title, normalized via the Softmax function. Raw sentiment scores are computed using weighted probability averaging, calculated as follows:

$$S_{raw} = \frac{1}{4} \times \sum_{i=0}^4 [i \times P(i)]$$

where S_{raw} denotes the raw sentiment score, ranging from [0, 1]; i represents the sentiment level index (0, 1, 2, 3, 4 correspond to one–five-star ratings, respectively); $P(i)$ indicates the probability predicted by the model that the title belongs to the i -th sentiment level; and the denominator 4 is used to normalize the score to the [0, 1] interval. Using this calculation method, a score of 0 indicates extremely negative sentiment, 1 indicates extremely positive sentiment, and values around 0.5 represent neutral sentiment.

Given that users typically exhibit heightened sensitivity to extreme emotions, this study further constructs an emotional intensity metric to quantify a headline's potential to evoke emotional responses. Extreme emotions are more effective than neutral emotions at capturing user attention and driving engagement. We calculate the emotional intensity score using the following conversion formula:

$$S_{intensity} = 2 \times |S_{raw} - 0.5|$$

where $S_{intensity}$ represents the emotional intensity score, ranging from [0, 1]. $|S_{raw} - 0.5|$ denotes the absolute deviation between the raw emotional score and the neutral point (0.5). This transformation function exhibits V-shaped symmetry: when S_{raw} approaches 0 (extremely negative) or 1 (extremely positive), $S_{intensity}$ approaches 1, indicating strong sentiment; when S_{raw} approaches 0.5 (neutral sentiment), $S_{intensity}$ approaches 0, indicating flat sentiment. This design ensures that extreme positive and negative sentiments receive equivalent intensity evaluations.

Based on the distribution characteristics of emotional intensity scores, this study categorizes headlines into two attractiveness classes: high-attractiveness headlines ($S_{intensity} \geq 0.6$) correspond to raw scores $S_{raw} \leq 0.2$ or $S_{raw} \geq 0.8$, indicating headlines with pronounced emotional bias; and low-attractiveness headlines ($S_{intensity} < 0.6$) correspond to raw scores $0.2 < S_{raw} < 0.8$, indicating relatively flat emotional expression.

3.6. Communication Effect Measurement

To conduct a comprehensive analysis, our study employed a dual approach to measure communication effectiveness. Our primary method involved analyzing four distinct dimensions of engagement as separate dependent variables to understand the nuanced impact of content features. As a supplementary method, we also constructed a single composite score for a holistic overview.

The four primary dependent variables were as follows:

- **Communication Breadth:** Measured by the total view count, representing the overall reach of the video. It is denoted by the letter B and serves as the foundational layer of video exposure.
- **Support Intensity:** A metric combining likes, coins, and favorites, reflecting audience approval and content value, denoted by the letter S

- **Discussion Activity:** A metric combining comments and shares, indicating deeper cognitive engagement, denoted by the letter *D*
- **Interactive Endorsement:** Measured by the total bullet comment count, capturing real-time audience participation, denoted by the letter *E*

For the supplementary analysis, these dimensions were combined into a holistic ‘Communication Effect’ (*C*) score. The weights assigned (*B*: 0.35; *S*: 0.25; *E*: 0.2; *D*: 0.2) were grounded in a hierarchical model of audience engagement, reflecting the different stages of the user funnel, ranging from awareness to advocacy.

Given that all count-based metrics exhibited strong positive skew, a natural logarithm transformation was applied to each of the four primary variables and the final composite score prior to regression analysis to normalize their distributions.

Our weighting scheme was based on a hierarchical model of audience engagement—a framework with a broad practical basis in digital media and marketing analytics. This model posited that different user actions represent distinct stages of an engagement funnel. Communication Breadth (*B*, View Count), as the foundation for all interaction, constituted the top of the funnel (Awareness), and was therefore assigned the highest weight (0.35). Support Intensity (*S*, Likes/Coins/Favorites) represented an initial level of positive user feedback and belonged to the middle layer (Consideration), thus receiving the next highest weight (0.25). Finally, Interactive Endorsement (*E*, Bullet Comments) and Discussion Activity (*D*, Comments/Shares) represented deeper levels of cognitive engagement and community interaction (Advocacy & Community) that typically resulted from the first two stages, and they were therefore assigned equal and slightly lower weights (0.2 each). This weighting structure reflected the logical progression and differential incidence rates of these user behaviors.

To validate the robustness of our findings against this specific weighting scheme, a sensitivity analysis was conducted. The regression analysis was repeated using two alternative schemes: (1) an equal weighting (0.25 for all components) and (2) an interaction-focused weighting (0.3 for *E* and *D*; 0.2 for *B* and *S*). This analysis confirmed that our core findings were highly stable. Specifically, the direction and statistical significance of the key predictor variables remained consistent across all three models. Furthermore, the model’s overall explanatory power, as indicated by the R-squared value, showed negligible variation (0.1215, 0.1219, and 0.1215, respectively). This demonstrated that our conclusions were robust and not an artifact of the chosen weighting configuration. Therefore, the comprehensive formula for our final communication effect is as follows:

$$C = \ln(0.35B + 0.25S + 0.2E + 0.2D)$$

The natural logarithm ($\ln$) transformation is applied to the weighted sum for critical statistical reasons. The distribution of raw communication effectiveness scores, like most social media engagement data, is highly positively skewed, with a long tail of high-performing videos. This skewness can violate the assumptions of ordinary least squares (OLS) regression, leading to unreliable parameter estimates. The log transformation corrects for this skew by making the distribution of the dependent variable more symmetric, which in turn helps to normalize the model’s residuals and stabilize their variance (i.e., address heteroscedasticity). This procedure ensures that the results of our subsequent regression analysis are more robust and statistically valid.

To control for potential confounding variables, this study also collected foundational video metadata, including total video duration (in seconds), publication date (categorized as “newly published” or “earlier published” based on a cutoff point of July 2023), title text length (character count), presence of exclamation marks or question marks in the title, pres-

ence of numbers in the title, and creator account follower count tier (800,000–2.3 million). This metadata was integrated into subsequent analyses as control variables to ensure more precise identification of the influence of human subjects, vehicles, and emotional features on dissemination outcomes.

3.7. Ethical Considerations

This study, involving large-scale data collection from a public social media platform and the application of facial expression recognition, was designed with careful consideration of key ethical principles. Our approach is grounded in a commitment to user privacy and the responsible use of computational methods.

First, all video data analyzed was sourced exclusively from publicly accessible content on the Bilibili platform. We did not access, collect, or analyze any private, restricted, or non-public user data. The content creators of the videos had, by choosing to publish on the platform, made their work available to a public audience.

Second, our analysis is strictly de-identified and focused on content in the aggregate. The facial expression recognition model was employed as a quantitative tool to generate statistical features for our regression models (e.g., the proportion of time a certain emotion was displayed). The objective was to understand patterns in video content, not to identify, monitor, or make assessments about the specific individuals appearing in the videos. No personal identifiable information (PII) of the on-screen individuals or the video creators was collected or stored.

Finally, our data scraping process was confined to the video content and its associated public engagement metrics (e.g., view counts, likes). We did not collect user-specific data, such as commenter identities or viewer profiles, thereby ensuring the privacy of the platform's user base. This study adheres to the ethical guidelines for internet research by focusing on publicly available data and ensuring that the analysis does not intrude upon the privacy or rights of individuals.

4. Related Work

The automotive video dissemination effectiveness prediction framework constructed in this study employs a hierarchical technical implementation strategy involving two core visual tasks: object segmentation and facial expression detection. For the segmentation task of people and vehicles, we adopt a mature segmentation model pre-trained using weights from the COCO dataset [10]. This choice is based on the COCO dataset's high-quality and large-scale data advantages in terms of segmentation annotations, with its pre-trained models already capable of providing sufficiently precise pixel-level segmentation results to meet our requirements for calculating the area ratios of people and vehicles. More importantly, the technology for such general segmentation tasks is relatively mature, reducing the necessity for redundant development. In contrast, facial expression detection for people, as the core technological innovation of this study, directly impacts the accuracy of video emotional feature extraction, which in turn affects the reliability of the propagation effect prediction model. While specialized facial expression recognition models have advanced considerably, general-purpose object detection frameworks (e.g., standard YOLO variants) still face critical limitations—such as insufficient sensitivity to subtle texture changes and lack of region-specific attention—when applied to fine-grained emotional analysis in real-world video content. Therefore, we specifically designed the FER-MA-YOLO model to achieve specialized optimization for facial expression detection tasks. Subsequent sections will systematically review the research progress on facial expression recognition, YOLO series object detection methods, and hybrid architecture design, providing the theoretical

foundation and performance comparison benchmarks for the technical innovations of the proposed FER-MA-YOLO model in expression detection tasks.

4.1. Facial Expression Recognition

Facial expression recognition (FER) is a key branch of computer vision and artificial intelligence that has undergone a significant transformation from classical machine learning to deep learning paradigms over the past decade. Facial expressions serve as the core medium for conveying human emotions, reflecting diverse emotional signals, social interaction intentions, and psychological connotations. This study is based on the FER framework [11], comprehensively reviews its evolutionary trajectory and technological innovations, particularly emphasizing how the cumulative occurrence frequency of facial expressions (such as anger, fear, disgust, happiness, sadness, surprise, and contempt) in video sequences shapes engagement metrics such as view counts, likes, shares, and discussion depth, thereby contributing theoretical insights and empirical evidence for digital media content strategies and emotion-driven communication models.

The initial stage of FER was heavily influenced by psychologists Ekman and Friesen's basic emotion framework [12], which categorized facial expressions into six basic types: anger, fear, disgust, happiness, sadness, and surprise. Later, contempt was added to form a seven-category system. Early methods relied on manually constructed feature extraction tools combined with standard machine learning classifiers. In terms of feature capture, geometric methods examine the spatial layout, distance fluctuations, and angular deformations of facial landmarks to track expression evolution; texture methods utilize local binary patterns (LBP) [13] and histograms of orientation (HOG) [14] to extract quantitative descriptions of skin texture; dynamic methods focus on analyzing facial muscle movements and temporal sequence patterns, making them particularly suitable for continuous expression tracking in video streams. In the classification phase, techniques such as support vector machines (SVM) [15], k-nearest neighbors (KNN) [16], and random forest [17] are commonly used for category division. While these foundational tools provide initial support for cumulative expression counting in videos, they struggle to deeply analyze the long-term correlation between expression frequency and overall engagement metrics.

Nevertheless, earlier methods revealed a series of inherent shortcomings. Artificial features struggle to comprehensively capture the subtle variations and diverse forms of expressions in videos, particularly overlooking the cumulative effects of expression persistence and frequency accumulation. They fail to robustly uncover the intrinsic connections between expression distribution and metrics such as viewing counts and appreciation numbers.

The emergence of deep learning has substantially advanced FER, particularly enhancing the efficiency of video expression processing. Convolutional neural networks (CNNs) leverage their exceptional feature self-learning and multi-layer abstraction capabilities to seamlessly construct a complete spectrum from basic edge details to advanced conceptual representations [18], significantly improving the accuracy and adaptability of FER, thereby more accurately quantifying the cumulative frequency of various expressions within videos. Recently, the design of novel CNN structures such as residual links, dense interconnections, and multilevel feature aggregation has further enhanced the model's expressive potential and optimization robustness, making it more suitable for facial expression statistics in long-duration videos and uncovering potential patterns between expression frequencies and metrics such as sharing scale and discussion depth.

With technological advancement, the integration of FER into attention modules marks a pivotal shift [19]. Through spatial and channel-wise optimization, the model dynamically locks onto core facial segments and dominant feature paths within video frames, effectively filtering out background noise and irrelevant disturbances. This enhances the precision

of expression frequency measurement and more clearly captures how expression patterns drive metrics such as appreciation counts. Simultaneously, the advancement of multimodal integration has expanded the scope of FER [20].

4.2. Object Detection Method Based on the YOLO Series

Object detection is a core task in computer vision [21,22], aimed at identifying and precisely locating objects in images. Its mainstream technical approach has evolved from a two-stage to a single-stage process. While the two-stage method centered on a Region-based Convolutional Neural Network (RCNN) offers superior accuracy, its “region proposal-classification regression” workflow makes it difficult to meet the demands of real-time applications [23]. To address this bottleneck, single-stage detectors represented by the YOLO (You Only Look Once) series emerged, unifying the detection problem into an end-to-end regression task. With its exceptional balance between speed and accuracy, it quickly became the mainstream paradigm in both academia and industry.

The evolution of the YOLO series can be seen as demonstrating the history of object detection technology, continuously pushing the boundaries of theory and practice. The first phase (v1–v3) was the period of establishing and refining the foundational framework. YOLOv1 [24] pioneered the concept of single-stage detection. To overcome the shortcomings of its initial design, YOLOv2 [25] introduced batch normalization to stabilize training and adopted the anchor box mechanism to improve localization accuracy. Subsequently, YOLOv3 [26] marked the maturation of the series. It adopted a deeper Darknet-53 network and incorporated residual connections to optimize gradient flow, while innovatively designing a multiscale prediction structure that significantly enhanced detection capabilities for objects of different sizes, becoming a widely adopted performance benchmark for years to come.

The second phase was the “comprehensive” and engineering optimization period represented by YOLOv4 [27] and YOLOv5. During this period, YOLO’s development entered a community-driven fast track. YOLOv4 systematically integrated a series of the best optimization strategies at the time, such as the CSP network structure and PANet [28] feature fusion, achieving a significant leap in performance. Meanwhile, YOLOv5 focused on engineering usability and deployment flexibility. Its clear PyTorch implementation and flexible model-scaling strategy greatly promoted the widespread adoption of YOLO in the industry.

The third phase was a period of paradigm shift and deep innovation toward anchor-free architecture. This series began to shift toward anchor-free design on a large scale, enhancing model performance and flexibility through the decoupling of detection heads and the SimOTA dynamic label allocation strategy. Subsequent versions such as YOLOv6 [29], YOLOv7 [30], and YOLOv8 have continued to evolve this design philosophy, driving continuous innovation in network architecture, training strategies, and multi-task support, consistently setting new industry standards. Recently, YOLOv9 [31] optimized the network information flow from an information theory perspective, while YOLOv10 [32] achieved breakthrough progress in end-to-end deployment efficiency by removing NMS post-processing.

However, despite the YOLO series’ tremendous success in general object detection, applying it directly to specific and specialized tasks, such as facial expression recognition (FER), presents unique challenges. FER not only requires accurate localization of faces, but, more critically, the identification of expressions formed by subtle muscle movements. This demands two core capabilities from the model: first, high sensitivity to subtle texture changes in local key regions (e.g., eye corners, mouth corners); second, full utilization of abstract semantic information embedded in depth features.

Existing research has begun exploring adaptive modifications to YOLOv10. For example, Qi et al. (2024) attempted to refine YOLOv11 in their proposed TPGR-YOLO [33]. To effectively capture detailed features; this study introduced a channel-spatial collaborative attention mechanism to focus on key regions and designed an enhanced path aggregation network (PANet) to better fuse multiscale feature information in dynamic scenes.

However, such approaches often focus on enhancing the model's general feature-extraction capabilities. Through an in-depth analysis of the technical architecture of existing YOLO-FER methods, we have identified systemic deficiencies across three core dimensions. First, regarding feature fusion strategies, many existing approaches, such as those described in UNet [34], predominantly employ linear concatenation to integrate multiscale features. This parallel fusion fails to establish semantic connections between features at different levels, resulting in a lack of organic coupling between shallow-level detail information and deep-level abstract features. This severely limits the model's overall understanding of complex facial expression patterns. Second, in attention mechanism design, current approaches predominantly employ single-dimensional attention calculations. They either focus solely on channel-wise feature weight allocation or concentrate exclusively on spatial-domain region selection. This lack of synergistic optimization between "feature selection" and "region localization" prevents the precise capture of critical facial expression information at the fine-grained level.

This deficiency serves as the starting point for our research. To address these issues, this paper adopts the performance-balanced YOLOv11 as its foundational architecture. Through the design of the DGF dense connection module and the MDAM multiscale dilated attention module, specifically tailored for facial expression features, YOLOv11 undergoes a deep reconstruction. This aims to adapt general object detection architectures for the specialized task of facial expression recognition, thereby improving recognition accuracy, detection speed, and robustness for small-target expressions.

4.3. Hybrid Architectures for Real-Time Detection

The evolution of feature fusion techniques reflects a trajectory of gradual refinement from coarse to fine-grained approaches. Early methods like Feature Pyramid Network (FPN) initially addressed the challenge of detecting objects of varying sizes by constructing a multiscale feature hierarchy. However, their simple element-wise addition fusion strategy suffered from significant semantic information loss. Although Path Aggregation Network (PANet) enhanced feature propagation through bidirectional information flow, it fails to fundamentally bridge the representational gap between shallow-level details and deep-level semantics. In recent years, researchers have explored more intelligent fusion architectures: BiFPN [35] introduced learnable feature weighting mechanisms to achieve adaptive multiscale feature integration; while NAS-FPN [36] leverages neural architecture search to automatically discover optimal fusion topologies.

The introduction of dense connections revolutionized feature fusion paradigms. DenseNet [37]'s design philosophy—"connecting every layer to all subsequent layers"—not only maximized feature reuse, but crucially established direct information pathways across layers, effectively mitigating gradient vanishing in deep networks. Its application in object detection spawned innovative architectures like DenseBox, significantly enhancing small-object detection performance. Concurrently, the rapid advancement of attention mechanisms revolutionized visual tasks. By mimicking the selective attention capabilities of the human visual system, attention enables deep models to adaptively focus on the most discriminative information.

Future technological trends exhibit distinct convergence and intelligence characteristics. On the one hand, feature fusion is evolving from simple linear combinations toward

deep learning-based adaptive fusion, with neural architecture search technologies increasingly maturing in automated feature fusion architecture design. On the other hand, attention mechanisms are advancing toward multidimensional coordination and lightweight deployment, where achieving real-time inference while maintaining high accuracy remains a critical challenge. Particularly in edge computing and mobile device deployment scenarios, demands for computational efficiency are becoming increasingly stringent.

5. The Proposed FER-MA-YOLO Model

YOLOv11 is the latest advancement in the YOLO series of object detection frameworks, integrating C3k2 blocks, SPPF fast modules, and parallel spatial attention C2PSA blocks, which collectively enhance feature extraction and detection performance. While transformer-based detectors offer strong accuracy in classification tasks, they are generally less optimized for real-time, edge-deployable object detection—a critical requirement when processing hundreds of thousands of frames across 968 automotive videos. In contrast, YOLO's single-stage, anchor-free (in recent versions) architecture provides an excellent speed–accuracy trade-off and significantly lower computational overhead. Our FER-MA-YOLO contains only 8.91 M parameters and achieves 164.7 FPS on a standard GPU, making it highly suitable for large-scale video analysis in computational communication research. Moreover, recent YOLO variants (v6–v11) already incorporate attention mechanisms and advanced feature fusion, narrowing the gap with transformer-based approaches while retaining inference efficiency. This series includes various variants ranging from lightweight to large-scale; this study selects YOLOv11n as the baseline model due to its status as the current state-of-the-art in the YOLO lineage.

However, our adaptation is not aimed at general object-detection improvement, but at constructing a reliable measurement instrument for the specific visual features hypothesized to drive communication effectiveness. This paper proposes the FER-MA-YOLO (Facial Expression Recognition-Multi Attention-YOLO) model, as shown in Figure 1. By introducing a multi-dimensional attention mechanism, it achieves precise detection and classification of facial expressions.

The core innovation of FER-MA-YOLO lies in the construction of two key feature enhancement modules. The DGF (Dense Growth Feature Fusion) module addresses redundancy issues in traditional feature concatenation through an improved dense connection strategy, enabling efficient multiscale feature fusion. The MDAM (Multiscale Dilated Attention Module) module employs multiscale dilated convolutions and a dual attention mechanism to comprehensively model features across spatial, channel, and scale dimensions, significantly enhancing the model's ability to perceive key features. The overall architecture adopts an end-to-end learning framework, seamlessly integrating improved feature extraction, multi-dimensional attention enhancement, and detection classification heads. The DGF and MDAM modules form complementary feature enhancement strategies, with the former focusing on optimizing feature fusion efficiency and the latter reinforcing attention modeling in the spatial and channel dimensions. This collaborative design enables FER-MA-YOLO to maintain stable detection performance in complex scenarios while balancing computational efficiency.

5.1. DGF (Dense Growth Feature Fusion)

In the network architecture design of YOLOv11, the neck module extensively employs serial operations combined with upsampling techniques, effectively integrating multilevel feature information generated by the backbone network. This mechanism fuses fine-grained spatial information from shallow layers with abstract semantic representations from deep layers, providing rich multiscale feature inputs for subsequent detection modules.

However, this direct channel-based concatenation strategy reveals significant design flaws. Its lack of intelligent feature-selection capability leads to indiscriminate fusion of features from different depths, resulting in massive accumulation of redundant information. The network struggles to distinguish effective feature components from noise interference, severely limiting the performance of the prediction module. When confronting challenging detection tasks (such as multi-object overlapping scenarios), the model's recognition accuracy drops significantly.

Based on this analysis, we propose the DGF (Dense Growth Feature) fusion scheme. This innovative design draws inspiration from the core principle of densely connected networks; extensively validated through experiments, this architecture achieves higher computational efficiency and fewer parameters while outperforming residual networks. The DGF fusion scheme will significantly address the shortcomings of traditional concatenation methods, enhancing the detection accuracy and computational efficiency of the entire network system through a more precise feature fusion mechanism.

As illustrated in Figure 3, the DGF fusion scheme progressively concatenates shallow high-resolution details with deep low-resolution semantic features across stages. Each DGF stage receives multiscale objectives from different stages, with each stage having a size of $\left(\frac{H}{2^n} \times \frac{W}{2^n} \times C_n\right)$. The fused features are processed through the DGF (fusion network) stage, composed of multiple dense blocks, before being passed to subsequent refinement modules. Specifically, the dense connection mechanism promotes feature reuse by having each dense block receive features from all preceding dense blocks as additional input. Thus, each dense block first performs channel adjustment using $\text{Conv}_{1 \times 1}$. To achieve efficient multiscale representation, we employ Deep Separable Convolution (DWConv) [38] as the sampling module. This module systematically reduces the spatial resolution of input features through a hierarchical progressive downsampling strategy. DWConv decomposes standard convolution into depthwise convolution and dot-product convolution: depthwise convolution performs spatial filtering independently on each input channel, while dot-product convolution captures cross-channel information via $\text{Conv}_{1 \times 1}$. This decomposition significantly reduces computational cost and parameter count. First, apply DWConv to the input feature $f_n^e \in \mathbb{R}^{\frac{H}{2^n} \times \frac{W}{2^n} \times C_n}$ for spatial downsampling, generating the output feature $f_{n+1}^{e'} \in \mathbb{R}^{\frac{H}{2^{n+1}} \times \frac{W}{2^{n+1}} \times C_{n+1}}$, where H and W , respectively, denote the height and width of the input image, and $n \in (0, 1, 2, 3)$. Subsequently, a new feature map is generated through the Deep Separable Convolution module and concatenated along the channel dimension with the input features, expanding the channel count at a constant growth rate g . The initial channel count is denoted as C_{in} . After passing through n fully connected blocks, the total channel count increases to $(C_{in} + n * g)$. To address channel expansion, $\text{Conv}_{1 \times 1}$ channels are added after the fully connected blocks to restore the C_{in} channel count. The DGF mixer enables feature reuse by establishing fully connected links between channels. This approach significantly enhances the integration and retention of high-level semantic features, greatly improving its applicability and performance in real-time object detection tasks.

Although the DGF mixer effectively increases feature reuse through full connectivity, its design has limitations: First, DGF primarily focuses on feature fusion in the channel dimension, with relatively insufficient modeling of spatial feature relationships. This may lead to loss of critical spatial contextual information when handling objects with complex spatial distributions. Second, while dense connections facilitate feature propagation, they lack an adaptive weighting mechanism for varying feature importance, preventing the network from dynamically focusing on critical features. Furthermore, when addressing multiscale object-detection tasks, the DGF's fixed growth rate strategy may fail to adequately adapt to the feature representation demands of objects at different scales, revealing limitations particularly in small-object detection and dense object scenarios.

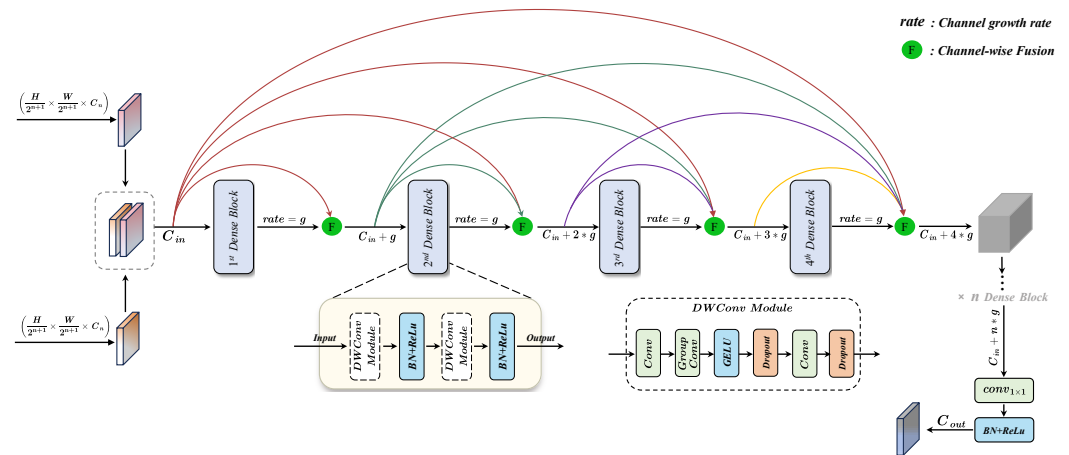


Figure 3. DGF network structure. (* is used as a symbol for multiplication).

5.2. MDAM (Multiscale Dilated Attention Module)

Dense connections, while promoting feature reuse, fail to effectively distinguish the importance of different features and struggle to fully model spatial feature relationships. Furthermore, while the feature maps output by the C2PSA module in the backbone undergo spatial attention processing, they can be further optimized through global channel information and refined spatial adjustments, forming a multi-level attention enhancement mechanism. To address this issue, we propose the multiscale Dilated Attention Module (MDAM) attention mechanism, whose network architecture is shown in Figure 4. Inspired by the dual-attention mechanism’s outstanding performance in the Convolutional Block Attention Module (CBAM) [39], our MDAM achieves comprehensive modeling of feature maps across three dimensions—spatial, channel, and scale—by constructing a multiscale receptive field and an adaptive weight distribution mechanism. In image processing tasks, objects can appear in varying sizes and shapes. Traditional convolutional layers typically possess fixed receptive fields, limiting their ability to capture features across different scales. By introducing parallel multiscale dilated convolutions (using varying dilation rates), we significantly expand the receptive field without resolution loss while capturing features at diverse scales. This enhances the model’s adaptability and recognition capabilities for multiscale features in images. Building upon this foundation, we incorporate channel attention and spatial attention to focus the model on critical regions within the image.

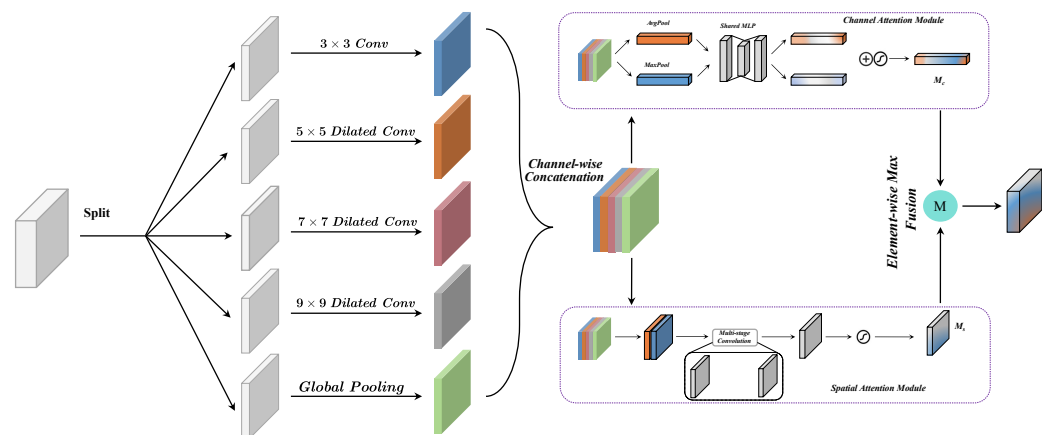


Figure 4. MDAM network structure.

Specifically, we adopt a parallel multiscale dilated convolution architecture. By setting different dilation rates $r \in (1, 3, 6, 9)$, we substantially expand the receptive field while

preserving feature map resolution. For the input feature map $x \in (\mathbb{R}^{H \times W \times C})$, the multiscale feature extraction and dual attention fusion process can be unified as follows:

$$F_{\text{cat}} = \text{Concat} \left[\text{Conv}_{1 \times 1}(X), \text{Conv}_{3 \times 3}^{(r_i)}(X), \text{Upsample}(\text{Conv}_{1 \times 1}(\text{GAP}(X))) \right]$$

where $\text{Conv}_{3 \times 3}^{(r_i)}$ denotes a 3×3 dilated convolution with an expansion rate of r_i , while GAP represents a global average pooling operation.

At the end of both attention modules, an element-wise max pooling operation integrates the outputs of the channel attention and spatial attention. Building upon the acquisition of multiscale features, the dual channel–spatial attention mechanism achieves feature weighting through a max fusion strategy, as expressed by the following formula:

$$F_{\text{output}} = \text{Conv}_{1 \times 1}(\max(F_{\text{cat}} \odot \sigma(\text{Conv}_{1 \times 1}(\text{GAP}(F_{\text{cat}}))), F_{\text{cat}} \odot \sigma(\text{Conv}_{1 \times 1}(F_{\text{cat}}))))$$

where $\odot$ denotes element-wise multiplication, σ represents the Sigmoid activation function, and the *max* operation achieves adaptive fusion of channel attention and spatial attention. We effectively combine multiscale features with attention mechanisms to ultimately output fused feature maps rich in multiscale information and precise attention weights. This not only addresses the DGF module's relative weakness in modeling spatial feature relationships, but also demonstrates through experimental performance that our proposed MDAM complements C2PSA to achieve synergistic feature enhancement effects.

6. Experiments

6.1. Methodological Justification for a High-Precision Feature Extractor

The use of the proposed FER-MA-YOLO model was methodologically critical for the subsequent communication analysis. The model's performance gains over the baseline—+5.72% in mAP50, +5.45% in mAP75, and +5.66% in mAP95—are considered highly significant in the field of object detection, where a 1–2% improvement is already deemed meaningful. This substantial increase in accuracy, particularly on stricter IoU thresholds (mAP75, mAP95), is essential for reliably capturing the subtle and often ambiguous signals of facial expressions.

The regression variables in this study (e.g., ratios of emotional expressions) are aggregates of thousands of frame-level classifications. Using a less precise feature extractor would introduce significant measurement error into these predictors. It is a well-established statistical principle that such errors can attenuate regression coefficients, increase the likelihood of Type II errors, and ultimately obscure the nuanced relationships this study seeks to uncover. Therefore, employing a feature extractor with state-of-the-art accuracy was a necessary step to ensure the validity and robustness of our findings, minimizing statistical noise and allowing for a more sensitive test of our research hypotheses. This principle holds even when comparing high-performing SOTA models, as small but consistent advantages in key metrics like precision can prevent the accumulation of thousands of misclassifications across a large dataset, thereby providing a more reliable foundation for statistical inference.

6.2. Experimental Settings

Datasets: We evaluated the facial expression recognition performance of the proposed FER-MA-YOLO model on the public “Facial Expression Recognition-9 Emotions (FER-9)” dataset (Rimi, 2024), available on the Kaggle platform. This dataset comprises 68,284 RGB images, each annotated with corresponding pixel-level emotion labels. To optimize training efficiency, we reduced the dataset size to 20% of the original data, totaling 13,657 images. The reduced dataset was split into a training set (10,926 images) and a validation set

(2731 images) at an 8:2 ratio. To enhance model generalization and mitigate overfitting risks, we employed data augmentation techniques, including random scaling ($\pm 15\%$ amplitude), horizontal flipping, and random cropping. Experimental design strictly adhered to the dataset partitioning and augmentation strategy to ensure the reliability and reproducibility of evaluation results.

Implementation Details: The model was trained on 8 GTX 1080Ti GPUs. To evaluate the effectiveness of the proposed FER-MA-YOLO model in facial expression detection, we conducted experiments on the “Facial Expression Recognition-9 Emotions (FER-9)” dataset. To enhance model generalization and reduce overfitting, we employed data augmentation techniques including random scaling ($\pm 15\%$), horizontal flipping (probability 0.5), random cropping, HSV adjustment (Hue: 0.015; Saturation: 0.7; Lightness: 0.4), translation (0.1), scaling (0.5), and random erasure (probability 0.4).

All baseline models (YOLOv5 through YOLOv11) were retrained from scratch on the same FER-9 dataset using identical data splits (10,926 training/2731 validation images), identical augmentation pipelines (random scaling, flipping, HSV jitter, etc.), and identical hyperparameters (120 epochs, batch size 200, input size 640×640 , initial learning rate 0.01, momentum 0.937, weight decay 0.0005, and warm-up for the first 3 epochs), ensuring that performance differences reflect true architectural distinctions rather than training protocol variations.

Evaluation Metrics: To comprehensively evaluate the performance of the FER-MA-YOLO model in facial expression detection tasks, we adopted a multi-dimensional evaluation metric system. This includes accuracy metrics (Precision, Recall, and F1-Score), mean average precision (mAP@50, mAP@75, and mAP@95), and model efficiency metrics (parameters Param (M), FLOPs (G), and inference speed FPS). These metrics were selected based on standard practices in the object detection domain to ensure the reliability of the results. Subsequent experimental results will compare the performance of FER-MA-YOLO with other baseline models based on these metrics.

- **Precision:** Precision measures the accuracy of a model’s predictions, representing the proportion of samples correctly classified as positive among all samples predicted as positive.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall:** The recall rate measures a model’s ability to identify true positives, representing the proportion of all genuine positive samples correctly recognized by the model.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-score:** The F1-score is the harmonic mean of Precision and Recall, providing a comprehensive evaluation of a model’s overall performance. It is particularly well-suited for handling datasets with imbalanced categories.

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP + FP + FN}$$

TP (True Positive) represents the number of target instances correctly detected by the model; FP (False Positive) represents the number of background areas or duplicate instances incorrectly detected by the model; FN (False Negative) represents the number of actual target instances missed by the model; and TN (True Negative) is a concept not typically used directly in object detection.

- **mAP (Mean Average Precision):** Average Precision (AP) is a measure of the Precision score along the Precision–Recall (PR) curve at different recall thresholds. mAP is

the average AP across all object categories. Specifically: mAP50 denotes the average average precision (AP) when the IoU threshold is set to 0.5; mAP75 denotes the average average precision (AP) when the IoU threshold is set to 0.75; mAP95 evaluates model performance across IoU thresholds from 0.5 to 0.95. It averages the mAP values corresponding to all IoU thresholds between 0.5 and 0.95. These three metrics collectively form a comprehensive evaluation framework, holistically reflecting the model's detection performance across different precision requirements.

6.3. Performance Comparisons

To evaluate the performance of the FER-MA-YOLO model, this section compares our proposed FER-MA-YOLO model with other YOLO-based models on the FER-9 dataset. Comprehensive evaluation metrics including Precision, Recall, F1-Score, mAP (mean average precision), model parameter count (Param(M)), floating-point operations per second (FLOPs(G)), inference time (Inference Time (MS)) and frames per second (FPS) were employed to measure the accuracy and efficiency of each model. The quantitative comparison results are summarized in Table 1.

Table 1. Quantitative performance comparison of FER-MA-YOLO and other models on the FER-9 dataset, where the best results are emphasized in bold.

Model	Precision ↑	Recall ↑	F1-Score ↑	mAP50 ↑	mAP75 ↑	mAP95 ↑	Param (M) ↓	FLOPs (G) ↓
YOLOv5_n	80.52	73.47	76.77	81.98	67.27	59.32	2.18	5.8
YOLOv5_s	80.34	78.72	79.51	84.26	68.87	62.16	7.81	18.7
YOLOv6_n	78.13	75.50	76.78	81.38	68.07	60.27	4.15	11.5
YOLOv6_s	80.53	74.63	77.45	83.81	69.97	62.07	15.97	42.8
YOLOv8_n	78.82	76.24	77.51	82.29	70.37	61.56	2.68	6.8
YOLOv8_s	80.49	77.36	78.88	84.22	71.24	63.69	9.82	23.4
YOLOv9_t	77.74	75.69	76.69	82.05	68.40	61.28	1.97	7.6
YOLOv10_n	77.69	69.99	73.61	79.66	66.11	58.05	2.26	6.5
YOLOv10_s	76.03	72.36	74.14	80.92	68.84	60.79	7.22	21.4
YOLOv11_n	72.05	75.06	73.50	79.86	66.43	58.47	2.58	6.3
YOLOv11_s	77.93	76.94	77.41	82.40	68.46	61.60	9.41	21.3
FER_MA_YOLO (Ours)	83.46	78.76	80.99	85.58	71.88	64.13	8.91	20.9

The FER-MA-YOLO model demonstrated outstanding performance across multiple key metrics, achieving 83.46% Precision, 78.76% Recall, and 80.99% F1-Score. Particularly for the mAP metric, FER-MA-YOLO demonstrated robust performance across different IoU thresholds (mAP50 = 85.58%, mAP75 = 71.88%, mAP95 = 64.13%), showcasing stability in high-precision detection and making it especially suitable for facial expression recognition tasks in complex environments. Quantitative experiments demonstrate that, compared to the novel algorithm YOLOv11, FER-MA-YOLO achieves improvements of 11.41% (Precision), 3.7% (Recall), and 7.49% (F1-score) over the n version, and 5.53% (Precision), 1.82% (Recall), and 3.58% (F1-score) over the s version on the FER-9 dataset, while reducing the number of parameters by 0.5M. Our model also demonstrates superior performance in mAP metrics compared to other benchmark models. To further assess practical deployability, we benchmarked inference latency across hardware platforms. An NVIDIA GTX 1080 Ti GPU, FER-MA-YOLO processed each frame at an average of 6.068 ms (≈ 164.7 FPS). On a CPU-only system (Intel(R) Xeon(R) CPU E5-2603 v4 @ 1.70 GHz), the latency increased to 288.06 ms per frame (≈ 3.47 FPS). The quantitative comparison results are summarized in Table 2. While this CPU speed is not suitable for real-time streaming, it remains viable for offline batch analysis of moderate-sized datasets. The model's computational load of 20.9 G FLOPs reflects a favorable balance between accuracy and efficiency, especially when

compared to alternatives like YOLOv8_s (23.4 G FLOPs) that achieve lower mAP50 (84.22% vs. our 85.58%). This efficiency profile makes FER-MA-YOLO well-suited for large-scale video analysis in both academic and industrial settings.

Table 2. Inference speed (CPU/GPU) comparison of FER-MA-YOLO and other models on the FER-9 dataset.

Model	Inference Time-CPU (ms) ↓	FPS (CPU) ↑	Inference Time-GPU (ms) ↓	FPS (GPU) ↑
YOLOv5_n	100.69	9.93	2.335	428.2
YOLOv5_s	205.66	4.86	4.559	220.4
YOLOv6_n	106.19	9.42	2.347	426.1
YOLOv6_s	301.42	3.32	4.981	200.7
YOLOv8_n	106.67	9.37	2.543	393.2
YOLOv8_s	220.88	4.53	5.073	197.1
YOLOv9_t	101.37	9.86	2.373	292.8
YOLOv10_n	120.68	8.29	2.846	351.3
YOLOv10_s	270.31	3.70	5.634	177.4
YOLOv11_n	235.53	4.25	2.615	382.3
YOLOv11_s	242.30	4.13	5.478	182.5
FER_MA_YOLO (Ours)	288.06	3.47	6.068	164.7

It is worth noting that while FER-MA-YOLO achieves state-of-the-art accuracy, its inference speed of 164.7 FPS remains highly competitive, particularly when compared to other high-accuracy models. For instance, YOLOv8_s, which has a similar parameter count (9.82 M vs. our 8.91 M), operates at 197.1 FPS. This efficiency is primarily attributed to two design choices. First, our model is built upon the inherently efficient YOLOv11_n baseline, which provides a strong foundation for real-time performance. Second, the Dense Growth Feature Fusion (DGF) module employs depthwise separable convolutions (DWConv) for its downsampling operations. As detailed in Section 5.1, DWConv significantly reduces computational cost and parameter count compared to standard convolutions, thereby mitigating the speed penalty typically associated with adding new architectural components. Consequently, FER-MA-YOLO strikes a favorable balance between high detection accuracy and real-time inference capability, making it well-suited for processing large-scale video datasets in computational communication research.

Comprehensive analysis of comparative results confirms that FER-MA-YOLO achieves industry-leading performance in facial expression recognition tasks. Figure 5 visually demonstrates our model’s performance on automotive videos, revealing results that closely resemble real-world scenarios. To further explore the advantages of FER-MA-YOLO, this study will conduct ablation experiments in subsequent sections to analyze the contributions of each model component.



Figure 5. Extracting features for communication analysis from video frames: Identifying emotion via FER-MA-YOLO (prediction boxes) and quantifying key subject presence (e.g., humans and cars as shown) via Instance Segmentation (masks), with (a–d) representing four distinct car videos, respectively.

6.4. Ablation Study

To validate the effectiveness of each key component in the FER-MA-YOLO model, we conducted systematic ablation experiments. Specifically, we used YOLOv11_n as the baseline model (BaseLine) and progressively added the Dense Growth Feature Fusion module (DGF) and multiscale Dilated Attention Module (MDAM) to analyze the specific contributions of each module to model performance. The ablation results are shown in Table 3.

Baseline Model Performance Analysis: The baseline model YOLOv11_n achieved 72.05% Precision, 75.06% Recall, and 73.50% F1-Score on the FER-9 dataset, with mAP50 at 79.86%. This baseline performance provides a solid reference for subsequent module improvements. Integrating the DGF module into the baseline model resulted in significant performance gains. Precision increased from 72.05% to 76.97% (+4.92%), F1-Score rose from 73.50% to 76.12% (+2.62%), and mAP50 improved from 79.86% to 80.21% (+0.35%). These results demonstrate that the DGF module effectively promotes multiscale feature fusion through an improved dense connection strategy, addressing redundancy issues in traditional feature concatenation. The addition of the DGF module increased the number of parameters from 2.58 M to 6.68 M and boosted FLOPs from 6.3 G to 19.2 G, reflecting the trade-off between performance gains and computational complexity. When integrating the MDAM module alone, Accuracy improved to 78.12% (+6.07%), Recall reached 77.12% (+2.06%), and F1-Score increased to 77.60% (+4.10%). Particularly on the mAP50 metric, the MDAM module elevated performance to 81.69% (+1.83%), indicating that the comprehensive modeling across spatial, channel, and scale dimensions through multiscale dilated convolutions and dual attention mechanisms significantly enhances the model's perception of key features. When both DGF and MDAM modules are integrated simultaneously, FER-MA-YOLO achieves optimal performance. The Accuracy reached 83.46% (+11.41% over baseline), Recall was 78.76% (+3.70%), and the F1-Score was 80.99% (+7.49%). On the critical mAP metrics, mAP50 improved to 85.58% (+5.72%), mAP75 reached 71.88% (+5.45%), and mAP95 was 64.13% (+5.66%). These results fully demonstrate the synergistic effect between the two modules: DGF focuses on optimizing feature fusion efficiency, while MDAM enhances attention modeling in both spatial and channel dimensions, forming complementary feature-enhancement strategies.

Table 3. Results of the ablation experiments, where the best results are emphasized in bold.

BaseLine	DGF	MDAM	Precision ↑	Recall ↑	F1-Score ↑	mAP50 ↑	mAP75 ↑	mAP95 ↑	Param (M) ↓	FLOPs (G) ↓
✓	✗	✗	72.05	75.06	73.50	79.86	66.43	58.47	2.58	6.3
✓	✓	✗	76.97	75.29	76.12	80.21	68.33	60.14	6.68	19.2
✓	✗	✓	78.12	77.12	77.60	81.69	68.45	60.42	4.81	8.1
✓	✓	✓	83.46	78.76	80.99	85.58	71.88	64.13	8.91	20.9

7. Analysis and Results

This section presents the results of our quantitative analysis, examining the impact of multimodal features on a suite of video dissemination effectiveness metrics, including Communication Breadth, Support Intensity, Discussion Activity, and Interactive Endorsement. These features were extracted using our proposed high-performance FER-MA-YOLO model and other quantitative methods detailed previously. It is important to note that our primary analytical goal was statistical inference—to understand the relationships between specific content features and engagement metrics—rather than predictive modeling. Therefore, our main regression results are based on the full sample of 968 videos.

The results of our multiple regression analyses are presented in Table 4. To provide a granular understanding of how content features influence engagement, we first detail the findings for the four distinct dependent variables. The final column reports the results for the supplementary composite ‘Communication Effect’ score, which offers a general summary of the overall trends.

Table 4. Regression analysis of video features and communication effects of car videos.

Variables	Communication Breadth	Support Intensity	Discussion Activity	Interactive Endorsement	Communication Effect
Happy Expressions Time Ratio	0.0556	0.111 *	−0.0617	0.1057 *	0.0726
Natural Expressions Time Ratio	0.1216 **	0.2812 ***	0.0956 *	−0.0368	0.1494 ***
Surprised Expressions Time Ratio	0.0546	0.0257	0.1262	0.0557	0.0465
People Time Ratio	0.3956 ***	0.9759 ***	0.0126	0.1304	0.4585 ***
Vehicle Time Ratio	−0.2198 ***	−0.2228	−0.1243	−0.1851 *	−0.2663 ***
People Area Ratio	0.4612 ***	0.8841 ***	0.0494	0.1792 **	0.5016 ***
Vehicle Area Ratio	0.1407 *	0.1587	0.1963 **	0.2834 **	0.1558 *
Title Emotional Information Entropy	0.1416	−0.0331	0.2406	0.7021 ***	0.2534
Image Brightness	0.0365	−1.1180 **	−0.7569 *	−0.4200	0.1251
	−0.0422	0.0729	0.1887	0.2692	−0.0942
R-squared Value	0.1152	0.2567	0.0340	0.0339	0.1215
Total R-squared	0.1061	0.2490	0.0240	0.0240	0.1125

Note: Asterisks (*) indicate statistical significance at the $p < 0.05$ level, double asterisks (**) at the $p < 0.01$ level, and triple asterisks (***) at the $p < 0.001$ level. All p -values are two-tailed. Coefficients represent the expected change in the log-transformed dependent variable for a one-unit increase in the predictor, holding all other variables constant.

All regression analyses in this study were conducted using Ordinary Least Squares (OLS) multiple linear regression, implemented via the statsmodels library in Python 3.8.20. The dependent variables—Communication Breadth, Support Intensity, Discussion Activity, Interactive Endorsement, and the composite Communication Effect—were log-transformed to address the severe positive skew typical of social media engagement data. To assess potential multicollinearity among our predictors, we calculated the Variance Inflation Factor (VIF) for each independent variable. All VIF values were found to be below 5, indicating that multicollinearity is not a significant issue in our models.

7.1. The Impact of Emotional Expression and Title Sentiment

Our analysis reveals that emotional cues articulated by on-screen individuals are significant predictors of audience engagement. Consistent with H1, the **Happy Expressions Time Ratio** positively influenced both Support Intensity ($\beta = 0.11, p < 0.05$) and Interactive Endorsement ($\beta = 0.11, p < 0.05$). The **Natural Expressions Time Ratio** emerged as a particularly strong and pervasive predictor, consistent with H2. It showed significant positive effects on Communication Breadth ($\beta = 0.12, p < 0.01$), Support Intensity ($\beta = 0.28, p < 0.001$), Discussion Activity ($\beta = 0.09, p < 0.05$), and the overall Communication Effect ($\beta = 0.15, p < 0.001$). In contrast, the **Surprised Expressions Time Ratio** showed no significant relationship with any performance metric, leading to the rejection of H3.

These regression findings are complemented by our exploratory analysis of emotional dosage. This analysis suggests a non-linear effect, where audience engagement tends to peak when happy emotions constitute 15–25% of the video duration, with a declining

trend observed when this proportion exceeds 30%. Furthermore, corroborating H8, the **emotional intensity of the title** was a powerful positive predictor. With a large coefficient ($\beta = 0.70, p < 0.001$), it was one of the strongest single predictors of Interactive Endorsement in our model, highlighting the critical importance of textual framing in stimulating user interaction.

7.2. The Impact of Visual Presence: People and Vehicles

The visual presence and prominence of core subjects—people and vehicles—demonstrated a nuanced influence on dissemination. As predicted in H4 and H5, human elements were crucial drivers of engagement. The **People Time Ratio** was a strong positive predictor of Communication Breadth ($\beta = 0.40, p < 0.001$), Support Intensity ($\beta = 0.98, p < 0.001$), and the overall Communication Effect ($\beta = 0.46, p < 0.001$). Similarly, the **People Area Ratio** significantly predicted Communication Breadth ($\beta = 0.46, p < 0.001$), Support Intensity ($\beta = 0.88, p < 0.001$), and Interactive Endorsement ($\beta = 0.18, p < 0.01$). These results underscore the vital role of a human presence in attracting and retaining viewer attention.

Interestingly, the presentation of vehicles showed contrasting effects depending on whether it was measured by duration or screen area. Supporting H7, a larger **Vehicle Area Ratio** positively influenced multiple metrics, including Communication Breadth ($\beta = 0.14, p < 0.05$), Discussion Activity ($\beta = 0.20, p < 0.01$), and Interactive Endorsement ($\beta = 0.28, p < 0.01$). However, a longer Vehicle Time Ratio had a significant negative impact on Communication Breadth ($\beta = -0.22, p < 0.001$), Interactive Endorsement ($\beta = -0.19, p < 0.05$), and the overall Communication Effect ($\beta = -0.27, p < 0.001$). This finding directly contradicts our initial hypothesis and leads to the rejection of H6. It suggests a “quality over quantity” principle is at play for vehicle presentation: prominent, detailed shots are beneficial, but excessive, uninterrupted screen time dedicated solely to the vehicle may diminish viewer engagement and social appeal.

7.3. The Influence of Visual Complexity and Brightness

Finally, we examined the role of overall visual aesthetics. As predicted in H9, Information Entropy, a measure of visual complexity, was found to have a significant and strong negative influence on audience support. It negatively predicted both Support Intensity ($\beta = -1.12, p < 0.01$) and Discussion Activity ($\beta = -0.76, p < 0.05$). The large magnitude of these coefficients suggests a substantial penalty for videos with excessive visual clutter or detail, which may overwhelm the audience and hinder their willingness to engage. In contrast, **Image Brightness** showed no significant relationship with any of the dissemination metrics, leading to the rejection of H10 and indicating that it is not a key factor in driving engagement in this context.

8. Discussion

8.1. Interpretation of Key Findings

This study systematically investigates the relationship between the multimodal characteristics of automotive video content and its dissemination effectiveness on the Bilibili platform. Our findings, drawn primarily from the disaggregated analysis of the four engagement dimensions, reveal a complex interplay of emotional, visual, and textual features. By examining these dimensions separately, we can uncover nuanced patterns that a single composite score would obscure. For instance, the features that drive Communication Breadth are not always the same as those that foster Discussion Activity. The optimal ranges identified for happy (15–25%) and natural (40–50%) emotional expressions are consistent with Emotional Contagion Theory, where moderate positive emotions enhance viewer resonance. For a high-involvement product like an automobile, the value of a

stable, natural emotional tone aligns with the Elaboration Likelihood Model (ELM), as it likely fosters a credible information environment that encourages viewers' central route processing of product information.

Regarding visual content, while we did not identify a single optimal ratio, the regression results (Table 4) indicate a nuanced relationship: the People Time Ratio positively predicts engagement, whereas the Vehicle Time Ratio has a negative effect. This suggests that, while the vehicle is the core product, the human element is crucial for maintaining narrative engagement and social appeal.

The rejection of *H6*—the finding that a longer Vehicle Time Ratio negatively impacts communication effectiveness—warrants a deeper theoretical explanation. This counterintuitive result can be interpreted through the complementary lenses of Cognitive Load Theory and Narrative Transportation Theory.

From a Cognitive Load perspective, while the vehicle is the core subject, extended segments focusing exclusively on it without human guidance can overwhelm the viewer. A human presenter acts as a crucial cognitive filter, directing the audience's attention to key features and contextualizing the information. Without this guidance, the video risks becoming a dense "visual catalog," imposing a high extraneous cognitive load on the viewer, who must independently decide what is important. This mental fatigue can lead to decreased attention and a lower propensity to engage further.

Simultaneously, from a Narrative Transportation standpoint, the human presenter is the primary vessel for the story. Viewers are transported not just by the product, but by the experience, emotion, and narrative woven by the on-screen persona. When the presenter disappears for prolonged periods, the narrative thread is broken. This disruption pulls the audience out of their immersive state of "transportation," shifting their role from that of an engaged story participant to a passive product observer. This break in immersion logically leads to reduced engagement, such as lower continued viewership (Communication Breadth) and fewer real-time interactions (Interactive Endorsement).

In synthesis, the human element is not merely decorative; it is a vital functional component that manages cognitive load and sustains narrative transportation. Our findings empirically demonstrate that overemphasizing the product at the expense of the storyteller is a communicatively ineffective strategy in this context.

8.2. Theoretical and Methodological Contributions

The primary theoretical contribution of this work is the establishment of a quantitative model linking multimodal content features to communication outcomes in a vertical domain. It moves beyond qualitative analysis by offering a data-driven framework, providing new empirical evidence for theories such as Framing Theory (via title sentiment) and Uses and Gratifications in a computational context.

Methodologically, a key contribution lies in the development of the FER-MA-YOLO model. By integrating the DGF and MDAM modules, this study provides a robust solution for fine-grained visual analysis tasks like facial expression recognition within complex video environments. This demonstrates the value of tailoring computer vision models for specific communication research questions, and opens avenues for applying similar specialized models to analyze other non-verbal cues in digital media.

8.3. Practical Implications

For automotive marketers and content creators, our findings offer several data-driven strategies. An effective emotional arc might be structured around an "appropriately cheerful + stable and natural" ratio. Visually, a "character-led + product-core" principle should be maintained, ensuring that human presence guides the narrative without excessively

long, uninterrupted segments focused solely on the vehicle. Finally, the significant impact of title sentiment underscores the need for consistency between the emotional framing of the title and the emotional cues within the video itself. The computational efficiency of our FER-MA-YOLO model—requiring only 6.068 ms per frame on a standard GPU—ensures that the proposed framework can be practically deployed for large-scale analysis of automotive video content, enabling timely, data-driven optimization of marketing strategies.

8.4. Limitations and Future Research

This study has several limitations. First, a limitation of our study is the modest R-squared values in our regression models (e.g., $R^2 = 0.0340$ for Discussion Activity), which indicates that the multimodal features included in our analysis explain a limited portion of the variance in communication effectiveness. This result objectively reflects the complexity of the phenomenon; beyond in-video features, engagement is co-determined by numerous unmeasured factors, including platform algorithms, the creator–fan relationship, and the broader social context of hot topics. Despite this, the core value of this study is explanatory. The statistical significance of several independent variables—such as People Time Ratio, Natural Expressions Time Ratio, and Vehicle Area Ratio—provides empirical proof that the design of the video content itself has a measurable and significant effect on communication outcomes. By isolating these relationships, our research provides a theoretical foundation for optimizing communication effectiveness from the content production side, moving beyond anecdotal advice to offer data-driven principles for content strategy. Second, our data is drawn exclusively from the Bilibili platform. While this platform was chosen for its unique engagement features (e.g., “bullet comments”) and its central role in the Chinese automotive market, this single-platform focus inherently limits the generalizability of our findings. The communication dynamics, user expectations, and algorithmic environments of other major platforms, such as the global long-form content on YouTube or the short-form, virality-driven ecosystem of TikTok, are substantially different. Therefore, while our findings provide a robust model for the Bilibili context, their cross-platform applicability remains untested. A crucial direction for future research is to validate, adapt, and compare our multimodal framework across these different platforms to build a more universal model of automotive video communication effectiveness. Future research should extend this framework to other platforms like YouTube or TikTok. This would primarily require a conceptual and methodological rework of the Communication Effectiveness Measurement Layer to adapt to the unique engagement grammars of each platform. In contrast, the core Feature Extraction Layer, including the FER-MA-YOLO and BERT models, is highly portable and can be applied directly to video content from any source, providing a solid foundation for future cross-platform comparative studies. Third, from a computer vision perspective, the FER-MA-YOLO model was evaluated on a curated dataset using only three dominant emotion categories (happy, neutral, surprised). Its robustness on more diverse, “in-the-wild” video content and for a broader emotional repertoire—such as Ekman’s seven basic emotions or subtle states like “confusion” or “curiosity”—remains to be tested. Future research could address these limitations by employing experimental designs to test causality, conducting cross-platform and cross-cultural comparative studies, and extending the facial expression recognition model to more nuanced emotional categories to investigate their specific roles in automotive video dissemination. Furthermore, the architectural contributions of this study—the DGF and MDAM modules—were implemented and validated within a CNN-based framework. While these modules target general challenges in feature representation, their portability is not universal. The DGF module, designed for multiscale feature fusion, could be directly integrated into the neck of hierarchical Vision Transformers (e.g., Swin, PVT), which

produce spatial feature maps at multiple resolutions analogous to CNN backbones. The MDAM module, which relies on parallel dilated convolutions, requires a genuine 2D spatial structure and is, therefore, only applicable in architectures that preserve such representations—specifically, hierarchical Vision Transformers and hybrid models that combine convolutional layers with self-attention (e.g., CvT, CoaT). Empirical investigations on these integrations constitute a promising avenue for future research to rigorously assess the generalizability of our approach.

Future research also should explore the temporal and effective alignment between title sentiment and on-screen emotional expressions—for instance, how congruence or contrast between pre-viewing textual framing and in-video emotional dynamics jointly shapes audience engagement. Finally, while our methodological focus was on developing a computationally efficient model for large-scale analysis, we acknowledge that the validation experiments did not include a direct comparison against the latest Transformer-based architectures. Such a benchmark would provide the most comprehensive assessment of the trade-offs between inference speed and accuracy. Therefore, a crucial direction for future research is to conduct a thorough comparative analysis of FER-MA-YOLO against leading Transformer-based FER models to create a more complete performance landscape, offering valuable guidance for researchers in selecting the optimal tool for their specific analytical needs.

Author Contributions: Writing—original draft, investigation, Y.H. and Z.F.; Writing—outline of the paper, conceptualization, review and editing, supervision, funding acquisition, Y.H.; Y.H. and Z.F. contributed equally to this paper; Review and editing—W.L. All authors have read and agreed to the published version of the manuscript.

Funding: The authors are grateful for the finance support from General Program of National Social Science Fund of China (No. 22BXW040), Shanghai Jiao Tong University Participation in Research Program (No. T200PRP42015) and Ministry of Education Industry-Academia Collaborative Education Program (No. 231104575131914).

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Curiel, J. Viral communication through social media: Analysis of its antecedents. *RLCS Rev. Lat. Comun. Soc.* **2017**, *72*, 69–86. [CrossRef]
2. Zhao, X.; Lynch, J.G., Jr.; Chen, Q. Reconsidering Baron and Kenny: Myths and truths about mediation analysis. *J. Consum. Res.* **2010**, *37*, 197–206. [CrossRef]
3. Izadifar, P. A Study of Multimodal Analysis of a Volkswagen Car Advertisement Video. Seminar Paper, Technische Universität Darmstadt, Darmstadt, Germany, 2023. Available online: https://www.researchgate.net/publication/379034442_A_Study_of_Multimodal_Analysis_of_a_Volkswagen_Car_Advertisement_Video_Seminar_Multimodal_Discourse (accessed on 21 October 2025).
4. Bogdan, U.; Borza, A. Big Data Analytics and Organizational Performance: A Meta-Analysis Study. *Manag. Econ. Rev.* **2019**, *2*, 147–162. [CrossRef]
5. Li, H.; Zhang, N. Computer Vision Models for Image Analysis in Advertising Research. *J. Advert.* **2024**, *53*, 771–790. [CrossRef]
6. Weber, M.S.; Monge, P. The flow of digital news in a network of sources, authorities, and hubs. *J. Commun.* **2011**, *61*, 1062–1081. [CrossRef]
7. Lipschultz, J.H. *Social Media Communication: Concepts, Practices, Data, Law and Ethics*, 4th ed.; Routledge: London, UK, 2023. [CrossRef]
8. Pearce, W.; Özçelik, S.M.; Greene, A.K.; Teeling, L.; Bansard, J.S.; Omena, J.J. Visual cross-platform analysis: Digital methods to research social media images. *Information, Commun. Soc.* **2018**, *23*, 181–198. [CrossRef]

9. Hayes, A.F.; Preacher, K.J.; Myers, T.A. Mediation and the estimation of indirect effects in political communication research. In *Sourcebook for Political Communication Research*, 1st ed.; Routledge: London, UK, 2010.
10. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common objects in context. In Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
11. Liu, S.; Deng, W. Deep facial expression recognition: A survey. *IEEE Trans. Affect. Comput.* **2022**, *13*, 1195–1215. [\[CrossRef\]](#)
12. Terven, J.; Córdova-Esparza, D.M.; Romero-González, J.A. A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 1680–1716. [\[CrossRef\]](#)
13. Ojala, T.; Pietikäinen, M.; Mäenpää, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987. [\[CrossRef\]](#)
14. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893.
15. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [\[CrossRef\]](#)
16. Cover, T.M.; Hart, P.E. Nearest Neighbor Pattern Classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27. [\[CrossRef\]](#)
17. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [\[CrossRef\]](#)
18. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
19. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010.
20. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Liu, Y. VMamba: Visual State Space Model. *arXiv* **2024**, arXiv:2401.10166.
21. Tang, H.; Li, Z.; Zhang, D.; He, S.; Tang, J. Divide-and-Conquer: Confluent Triple-Flow Network for RGB-T Salient Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 1958–1974. [\[CrossRef\]](#)
22. Kang, S.; Hu, Z.; Liu, L.; Zhang, K.; Cao, Z. Object detection YOLO algorithms and their industrial applications: Overview and comparative analysis. *Electronics* **2025**, *14*, 1104. [\[CrossRef\]](#)
23. Yu, J.; Wei, Z.; Cai, Z.; Zhao, G.; Zhang, Z.; Wang, Y.; Xie, G.; Zhu, J.; Zhu, W.; Liu, Q. Exploring Facial Expression Recognition through Semi-Supervised Pre-training and Temporal Modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–18 June 2024; pp. 4880–4887.
24. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788. [\[CrossRef\]](#)
25. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525. [\[CrossRef\]](#)
26. Redmon, J.; Farhadi, A. YOLOv3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767. [\[CrossRef\]](#)
27. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934. [\[CrossRef\]](#)
28. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
29. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976. [\[CrossRef\]](#)
30. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 11832–11841. [\[CrossRef\]](#)
31. Wang, C.Y.; Liao, H.Y.M.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. *arXiv* **2024**, arXiv:2402.13616. [\[CrossRef\]](#)
32. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOv10: Real-Time End-to-End Object Detection. *arXiv* **2024**, arXiv:2405.14458.
33. Qi, X.; Xu, C.; Lu, Y.; Han, N.; Liu, H. TPGR-YOLO: Improving the traffic police gesture recognition method of YOLOv11. *Int. J. Adv. Comput. Sci. Appl. (IJACSA)* **2025**, *16*, 0160243. [\[CrossRef\]](#)
34. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241. [\[CrossRef\]](#)
35. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and efficient object detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 1540–1549. [\[CrossRef\]](#)
36. Ghiasi, G.; Lin, T.Y.; Le, Q.V.; Dollár, P. NAS-FPN: Learning scalable feature pyramid architecture for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019. [\[CrossRef\]](#)

37. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269. [[CrossRef](#)]
38. Chollet F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017. [[CrossRef](#)]
39. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.