HotComment: A Benchmark for Evaluating Popularity of Online Comments

Yafeng Wu
Huazhong University of Science and Technology
Wuhan, China

Yunyao Zhang
Huazhong University of Science and Technology
Wuhan, China

Liliang Ye
Huazhong University of Science and Technology
Wuhan, China

Guiyi Zeng
Huazhong University of Science and Technology
Wuhan, China

Junqing Yu
Huazhong University of Science and Technology
Wuhan, China

Chen Xu
Beijing Institute of Computer Technology and Applications
Beijing, China

Zikai Song*
Huazhong University of Science and Technology
Wuhan, China

Abstract

Online comments play a crucial role in shaping public sentiment and opinion dynamics on social media. However, evaluating their popularity remains challenging, not only because it depends on linguistic quality, originality, and emotional resonance, but also because stylistic preferences vary widely across platforms and user groups, causing the same comment to resonate differently in different communities. In this work, we present **HotComment**, a multimodal benchmark integrating video and text modalities that comprehensively quantifies popularity from three enhanced aspects: (1) Content Quality, which evaluates semantic similarity with ground-truth human comments and extends quality assessment through four interpretable dimensions; (2) Popularity Prediction, based on trends from models trained on real-world interaction data; and (3) User Behavior Simulation, which models the distribution of platform users and approximates **engagement scores** through an agent-based framework. Furthermore, we propose **StyleCmt**, inspired by **social ripple effects**, where **multiple stylistic dimensions align** to amplify socially resonant expressions and suppress incongruent ones.

CCS Concepts

• **Human-centered computing** → **Social media**; • **Computing methodologies** → *Natural language generation*; Machine learning; Modeling and simulation.

Keywords

Comment Generation, Social Media Analysis, Multimodal Dataset, Large Language Models

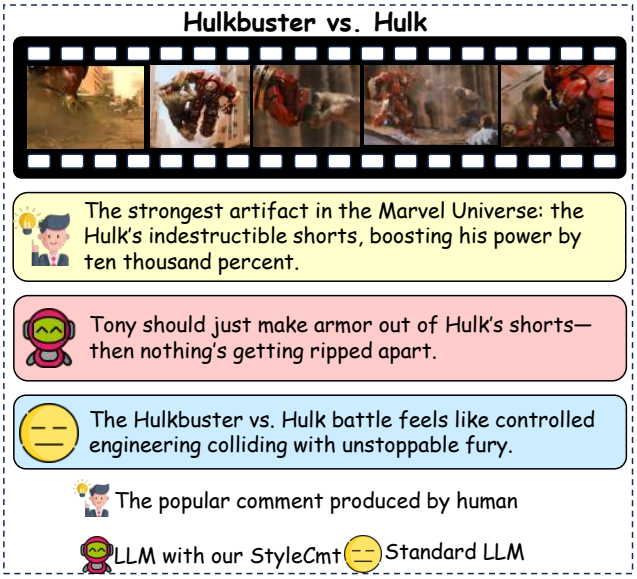

Figure 1: Example from the HotComment benchmark. Illustration of three types of comments for a given video: a real human-generated popular comment from the online platform, a StyleCmt-enhanced LLM-generated comment, and a standard LLM-generated comment without stylistic guidance.

1 Introduction

Online comments play a central role in shaping discourse on social media platforms[17, 79, 89]. With the rapid advancement of Artificial Intelligence and multimodal technologies [13, 28, 44, 46, 63], the landscape of popular content generation on social media has shifted from being purely human-driven to increasingly dominated by AI-generated content [87, 90]. However, evaluating their popularity remains challenging [7]. Popular comments are not determined by content relevance alone, but also by stylistic expression[35], context-aware wording[51], and the social pathways through which audiences encounter and react to content[1, 2].

Existing evaluation frameworks for social media comments mainly follow two directions. One line relies on lexical or semantic similarity metrics, such as BLEU and BERTScore[57, 85], to measure closeness to human references. Another extends evaluation from stylistic perspectives, incorporating factors such as humor, rhetorical devices, and creativity to better approximate human preference[9, 11, 93]. Recent benchmark efforts further emphasize multidimensional stylistic quality in comment-related generation settings[35]. While these methods capture important aspects of comment quality, they still provide only a partial account of popularity. Similarity-based metrics mainly reflect textual overlap or semantic closeness,

*Corresponding author.<skyesong@hust.edu.cn>

Benchmark	Multi	Scale		Content Quality		CrossAud.	
	-modal	Vis.	Txt.	LECIER	SCI	-plat.	Var.
TalkFunny[11]	✗	–	4k	✓ ✗ ✓	✗	✗	✗
Chumor 2.0[24]	✗	–	3k	✓ ✗ ✓	✗	✗	✗
Puns[73]	✗	–	2k	✓ ✓ ✗	✗	✗	✗
Oogiri-GO[93]	✓	100k	30k	✓ ✓ ✗	✗	✗	✗
NYT-Captions[26]	✗	3k	–	✗ ✗ ✗	✗	✗	✗
ViCo[69]	✓	20k	–	✓ ✗ ✓	✗	✗	✗
HOTVCOM[10]	✓	93k	–	✓ ✓ ✓	✓	✗	✗
GODBench[35]	✓	67k	–	✓ ✓ ✓	✓	✗	✗
HotComment	✓	34k	47k	✓ ✓ ✓	✓	✓	✓

Table 1: Benchmark comparison. Vis. and Txt. denote *Visual* and *Text*. LE, CI, ER, and SCI denote the four stylistic dimensions of *Content Quality*. Cross-plat. indicates datasets supporting cross-platform evaluation, and Aud. Var. denotes those considering audience variation.

whereas stylistic indicators are often treated as intrinsic properties of the comment itself. In practice, however, socially valued comments also depend on reasoning quality, constructiveness, and audience-sensitive engagement cues[19, 23, 32, 61]. Moreover, audience preferences vary across user groups, and exposure itself is shaped by demographic and cultural differences[20, 34, 62]. As a result, treating popularity as a universal textual property makes it difficult for existing frameworks to capture real engagement dynamics and the heterogeneous mechanisms underlying comment popularity[21, 71].

To address these challenges, we propose **HotComment**, a benchmark for evaluating the popularity of online comments. HotComment introduces a novel three-dimensional evaluation framework: (1) **Content Quality**, which evaluates semantic similarity with ground-truth human comments and additionally incorporates four stylistic dimensions ([Linguistic Expression], [Creative Imagination], [Emotional Resonance], and [Social and Cultural Influence]) to complement overlap-based metrics and capture how linguistic artistry, creativity, affective depth, and cultural propagation jointly relate to popularity; (2) **Popularity Prediction**, based on scores from engagement prediction models trained on large-scale real-world interaction data; and (3) **User Behavior Simulation**, which models heterogeneous user preferences through agent-based simulations of **engagement scores**. This multi-perspective design offers a more realistic and interpretable assessment of whether generated comments possess true popularity potential, establishing a robust foundation for future research on popularity-aware generation.

Inspired by the *Wave Interference Model*[25] and the *Uses and Gratifications Theory*[30], we propose **StyleCmt**, a novel framework that models the interaction among stylistic elements in linguistic space based on the principles of constructive and destructive interference. StyleCmt captures how different expressive patterns combine to amplify socially resonant forms while attenuating less compatible ones, generating comments that align with collective audience preferences. Experimental results demonstrate that this framework enables models to produce comments that more closely reflect human preference tendencies.

Our contributions are summarized as follows:

- We introduce a new large-scale and comprehensive online comment dataset with a multidimensional evaluation framework for online comments tasks.
- We propose StyleCmt, which simulates the interactions among different stylistic patterns, thus enabling the model to produce comments that resonate with dominant user preferences.
- Extensive experiments on the HotComment benchmark demonstrate that StyleCmt effectively captures stylistic and social dynamics, significantly improving the realism and engagement alignment of generated comments compared with baselines.

2 Related Work

2.1 Evaluation of comment quality

Early research on comment evaluation mainly focused on intrinsic textual quality, emphasizing coherence, fluency, and semantic consistency [12, 27, 84] with human-written references. Lexical and semantic metrics such as BLEU [57], ROUGE [48], and BERTScore [85] have been widely adopted to measure textual overlap and meaning alignment. Subsequent studies extended this line of work by incorporating stylistic and rhetorical dimensions into comment assessment. Existing research has examined humor [9, 11], irony [50], creativity [93], puns [68, 73], emotional expressiveness, and broader multidimensional stylistic quality [35]. These efforts deepen the understanding of how rhetorical and expressive features contribute to perceived comment quality. Beyond intrinsic quality and stylistic sophistication, another work has studied thoughtful and constructive comments as a distinct form of socially valued response. These studies show that constructiveness, reasoning quality, and conversational usefulness are not always equivalent to raw popularity feedback such as likes or replies, and should often be modeled separately [19, 23, 31, 32, 61]. Studies on constructive comments show that constructiveness, reasoning quality, and conversational usefulness often diverge from raw popularity signals like likes or replies, and should be modeled separately [19, 23, 31, 32, 61]. A similar mismatch occurs in image retrieval, where superficial engagement cues do not reliably reflect true relevance or user utility [18, 43, 45, 47]. These converging findings suggest that valuable feedback must often be disentangled from coarse popularity metrics.

Overall, existing comment evaluation methods mainly focus on intrinsic textual and stylistic quality, providing limited support for modeling platform-dependent engagement patterns and heterogeneous audience preferences. To address this gap, **HotComment** extends evaluation beyond the comment itself by introducing two complementary dimensions: **popularity prediction** and **user behavior simulation**.

2.2 Comment generation

Early studies on automatic comment generation [53, 59, 92] mainly relied on deep learning frameworks based on attention mechanisms [39, 78], structured modeling [67, 86, 91], or unsupervised matching between articles and comments. Later approaches improved diversity and structural representation through graph-based and retrieval-augmented designs [37, 76, 77], but they remained limited in semantic depth and rhetorical control. With the rise of large language models [4–6, 38], comment generation has advanced

substantially in multi-modal settings [10, 41, 42, 69], include both both text and visual cues [40, 64–66]. Recent studies have explored personalized comment generation [49, 81], popularity-aware social response generation [80], and controllable style steering during inference [83], showing that user identity, stylistic preference, and anticipated audience response are important conditioning factors. Existing methods, however, typically improve only a specific aspect of generation, such as creativity [93], humor [11], or contextual relevance [69]. In the multimodal setting, increasingly realistic tasks and datasets, such as LiveBot [52], PLVCG [82], knowledge-enhanced live video comment generation [8], and MMLSCU [54], have further enriched the problem setting. Nevertheless, most existing methods still treat controllable factors such as style, humor, or preference as isolated attributes, rather than explicitly modeling how multiple stylistic tendencies interact within a comment community and jointly shape audience resonance. To address this limitation, we propose **StyleCmt**, a cross-platform and multimodal framework that models comment-section preferences and generates comments aligned with dominant audience styles and expectations.

3 Challenge and Motivation

Challenge. Evaluating the popularity of online comments is more challenging than assessing intrinsic text quality alone. In social media, popularity is not a fixed property of the comment itself, but a conditional outcome jointly shaped by the comment, the source content, the platform context, and the exposed audience:

$$\pi(c \mid x, p, a) \tag{1}$$

where c denotes the comment, x the source content, p the platform context, and a the audience state. This formulation makes explicit that popularity depends on both contextual conditions and audience structure.

However, **existing benchmarks** and evaluation protocols mainly focus on semantic similarity, fluency, or isolated stylistic aspects, with limited ability to capture how these factors interact with real engagement mechanisms. A single popularity predictor reduces dissemination to a **coarse platform-level estimate**, while **generic judge-style evaluation** ignores exposure conditions, audience heterogeneity, and variation in user responses. Recent in-the-wild evidence from public human–LLM interactions on social media further shows that engagement in multi-party environments is highly asymmetric and socially embedded, making popularity inseparable from exposure context and audience structure [55]. As a result, current evaluation settings remain insufficient to determine whether a generated comment truly has real-world popularity potential.

Motivation and Design Rationale. These limitations motivate a benchmark design that approximates popularity from multiple complementary perspectives:

- **Content Quality** evaluates whether a generated comment is semantically appropriate and stylistically expressive, providing the textual foundation of popularity.
- **Popularity Prediction** captures platform-level interaction trends learned from real-world engagement data, reflecting aggregate platform tendencies.
- **User Behavior Simulation** models heterogeneous audience responses at the user level. Rather than acting as a generic LLM judge, it serves as a **data-grounded audience modeling component**. User profiles and agent weights are allocated according to real-world demographic statistics, regional internet population characteristics, and platform-specific user category distributions, ensuring that simulated reactions are conditioned on realistic audience composition.

Together, these three dimensions provide a structured approximation of popularity under heterogeneous exposure conditions, avoiding the reduction of complex popularity mechanisms to a single score and yielding a more faithful evaluation framework for socially grounded comment generation.

4 HotComment Benchmark

We first compare HotComment with the previous benchmark in Tab.1. Then, we introduce the three core dimensions of **HotComment**: **Content Quality**, which assesses linguistic and stylistic express; **Popularity Prediction**, which estimates engagement likelihood based on real-world interaction data; and **User Behavior Simulation**, which models audience preferences and interaction patterns to reflect realistic social dynamics.

4.1 Dataset Construction

Data Sources. We construct the **HotComment** dataset by collecting large-scale article–comment and video–comment pairs from multiple mainstream online platforms, including NetEase News, Tencent News, and Bilibili. These sources cover both textual and audiovisual content, enabling the study of multimodal communicative behavior and stylistic variation across platforms.

Scale and Composition. The dataset comprises **over 43,000 online articles** and **34,000 videos**, encompassing approximately **1.4 million content-comment pairs** in total. Each article or video is paired with corresponding user comments that reflect diverse linguistic styles and engagement behaviors. For each item, we collect both **popular comments** with high user engagement and **non-popular comments** with relatively lower engagement; the average number of comments per item varies by content type and platform characteristics.

Popularity Labeling. High-quality (popular) comments are defined as those ranked within the **top-15 by likes**, with like counts exceeding either **10% of the top comment** or an absolute threshold of **2,000 likes**. For low-quality (non-popular) comments, we select those posted on the *same day* as the content with like counts not exceeding **10**, collecting at most 10 such comments per item. This hybrid criterion ensures balanced representation while mitigating temporal bias.

4.2 Evaluation Methods

4.2.1 Content Quality. To comprehensively assess the intrinsic quality of generated comments, we evaluate not only their **semantic similarity** to human-written references but also their **stylistic expressiveness**[35] across four complementary dimensions:

- **Linguistic Expression** evaluates the rhetorical and writing artistry of comments, focusing on linguistic creativity such as humor, irony, metaphor, rhythm, and aesthetic fluency that enhance expressiveness and readability.

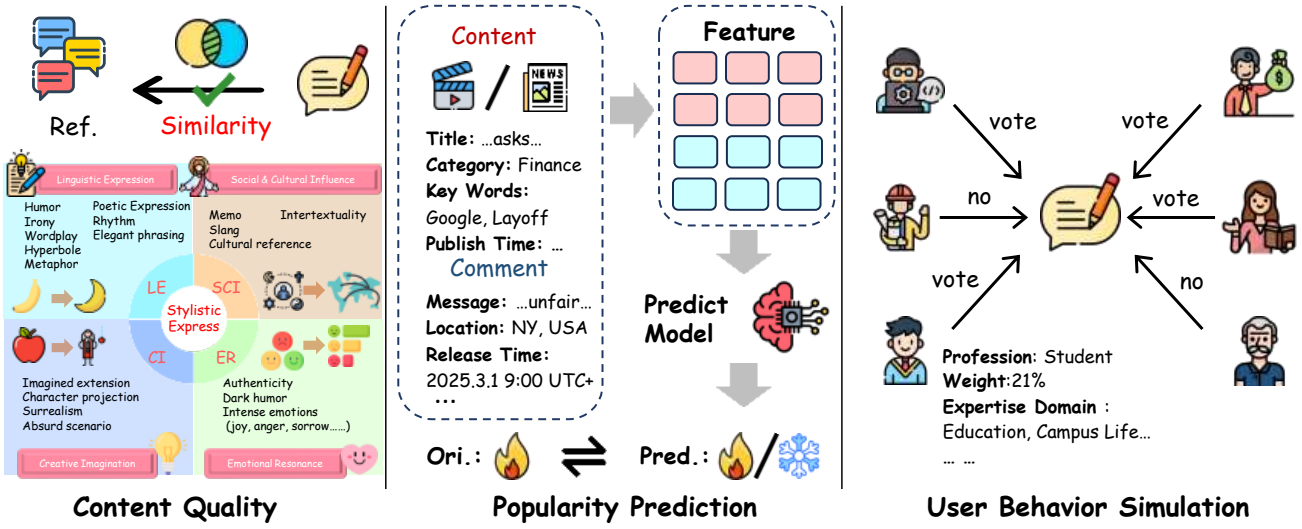

Figure 2: HotComment evaluates models from three key aspects:(1) Content Quality, assessed through multi-dimensional semantic comparison with top-k popular comments in the dataset; (2) Popularity Prediction, based on trends from popularity prediction models trained on real-world interaction data; (3) User Behavior Simulation, conducted via agent-based modeling of online user behavior such as thumbs-up.

- **Creative Imagination** measures the degree of originality and associative thinking within comments, capturing the ability to connect distant concepts or construct unexpected, imaginative scenarios that extend semantic boundaries.
- **Emotional Resonance** examines the emotional depth and attitudinal stance of a comment, emphasizing its capacity to evoke empathy, convey genuine sentiment, or express strong affective tones that engage readers.
- **Social and Cultural Influence** assesses the comment's potential for social propagation, including the use of memes, cultural references, and intertextual expressions that facilitate sharing, imitation, and collective resonance across communities.

4.2.2 Popularity Prediction. This dimension models **platform-level engagement preferences** to estimate how likely a comment would become popular within a given social media environment. Different platforms exhibit distinct user cultures and interaction mechanisms, leading to diverse definitions of "popularity". To reflect these differences, we train an individual **prediction model** for each platform using a fine-tuned BERT encoder with a task-specific MLP head, with real interaction data converted into binary popularity labels as supervision signals. Each model jointly encodes the contextual information of the article or video and the associated comment, and outputs a predicted popularity score that reflects engagement likelihood. Recent benchmark construction for social-media popularity prediction has also begun to emphasize temporal alignment and temporal dynamics, suggesting that engagement modeling should consider not only content–comment matching but also time-sensitive propagation patterns [72].

Training is performed under a binary classification objective, where popular and non-popular comments serve as positive and negative samples, respectively. We employ a combination of *cross-entropy loss* to ensure accurate classification and a *supervised contrastive loss* to enhance representation discrimination between high- and low-engagement comments. This setup enables the model to learn platform-specific engagement patterns while providing a reliable data-driven metric for evaluating the real-world popularity potential of generated comments. This design is broadly consistent with prior studies on popularity prediction and diffusion modeling, which treat engagement as a function of structural spread patterns, interaction dynamics, or jointly encoded content signals [3, 14, 22].

4.2.3 User Behavior Simulation. **User Behavior Simulation** models the composition of the exposed audience to approximate how comments would be perceived by different user groups within a realistic social environment. We formulate the exposure process as a **two-level hierarchical simulation**.

- At the top level, we classify potential viewers into two primary categories: **interested users** and **casual viewers**. For each platform–domain pair, we first assign an **Exposure Specificity Index (ESI)** that represents the baseline exclusivity of audience exposure. A **domain-specialized agent** analyzes the content to determine audience exposure patterns based on five key determinants of selective exposure: **channel verticality, distribution channel characteristics, event salience, emotional arousal level,** and **celebrity or authority involvement**. The agent outputs an adjusted proportion, denoted as p_I^*, representing the estimated share of interested users in the total audience. This modeling approach aims to reproduce the realistic composition of audiences who *actually see* the content, aligning with established theories of selective exposure, incidental news contact, and high-arousal dissemination.

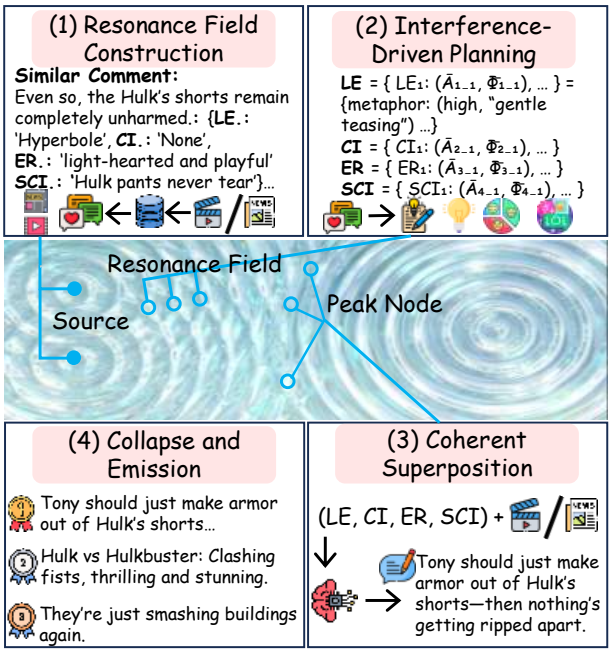

Figure 3: Overview of the StyleCmt Framework. The comment generation process is modeled as a form of wave interference in a stylistic field. The pipeline illustrates four consecutive stages: (1) *Resonance Field Construction.* Retrieving similar hot comments and decomposing their stylistic components across four dimensions (LE, CI, ER, SCI); (2) *Interference-Driven Planning.* Aggregating and identifying dominant stylistic patterns to form an interference blueprint; (3) *Coherent Superposition.* Generating multiple linguistic realizations under the same stylistic configuration; and (4) *Collapse and Emission* – selecting and refining the most resonant comment as the final output.

- At the lower level, we derive the distribution of user subgroups from *publicly accessible regional internet demographics* and *platform-specific user category statistics*, rather than preset rules. These data act as *data-driven priors* to construct heterogeneous audience segments with distinct interaction tendencies. This hierarchical design enables a user-centered interpretation of comment popularity beyond aggregate platform metrics.

Through this hierarchical design, the simulation offers a population-level perspective on how comments are likely to be received, thereby complementing platform-level popularity prediction with **user-centric interpretability**. Our formulation is also related to recent LLM-agent simulation frameworks, which emphasize role-conditioned behavior, social interaction, and aggregate engagement dynamics in synthetic but data-grounded online environments [36, 58, 60, 70]. This perspective is further supported by recent multi-agent social-media simulation work, which models public-opinion evolution through cognitively grounded agents and dynamic interaction environments [88].

5 StyleCmt Framework

5.1 Motivation and Computational Grounding

The **StyleCmt** framework is inspired by the way ideas and expressions spread through social interaction and gradually form shared preferences within a community. In our framework, historical comments under the same content context collectively define a **stylistic resonance field**, which captures the dominant expressive tendencies of the comment section.

To model this process, we represent each stylistic dimension as a vector with two components: **intensity**, which reflects the strength of community preference for that style, and **orientation**, which encodes its semantic and affective direction. When the stylistic representation of a generated comment is consistent with dominant historical patterns, the comment exhibits stronger stylistic coherence and better matches community preference. When the alignment is weak, the generated expression is less coherent with the surrounding discourse. This formulation provides a principled way to model stylistic preference in vector space and supports the generation of comments that are more consistent with collective audience expectations.

5.2 Framework Overview and Modeling Process

Resonance Field Construction. This step performs **contextual style distribution modeling**. To capture how a community tends to express itself, we construct the stylistic resonance field based on historical comments. Each comment is decomposed into four stylistic dimensions: Linguistic Expression, Creative Imagination, Emotional Resonance, and Social/Cultural Influence. For each dimension i, we estimate two properties from the data: (1) an *intensity* scalar $\bar{A}_i(\mathbf{x})$, representing how strongly the community prefers this stylistic pattern at context $\mathbf{x}$, and (2) an *orientation* unit vector $\bar{\mathbf{d}}_i(\mathbf{x})$, describing its typical stylistic inclination. We summarize the overall contextual stylistic tendency as a weighted vector aggregation:

$$\Psi_0(\mathbf{x}) = \sum_{i=1}^{4} \bar{A}_i(\mathbf{x})\, \bar{\mathbf{d}}_i(\mathbf{x}), \tag{2}$$

which acts as a compact mathematical representation of the community's baseline preference.

Interference-Driven Planning. This phase executes **alignment-driven style planning**. Given the base preference Ψ_0, we determine how a newly generated comment should adjust its stylistic mix. A candidate comment introduces a set of controllable stylistic contributions:

$$\mathbf{v}_i = A_i\, \mathbf{d}_i, \tag{3}$$

where A_i denotes the intended intensity of stylistic pattern i, and $\mathbf{d}_i$ describes its stylistic direction. To evaluate how well this candidate aligns with community tendencies, we compute the *interaction score* via the dot product of their orientation vectors (equivalent to cosine similarity):

$$I_{ij} = \mathbf{d}_i \cdot \bar{\mathbf{d}}_j. \tag{4}$$

A positive value indicates that the intended direction $\mathbf{d}_i$ is compatible with the community inclination $\bar{\mathbf{d}}_j$, conceptually mimicking constructive interference. The planning objective selects $\{A_i, \mathbf{d}_i\}$

that maximize the overall alignment:

$$\mathcal{E}_{\text{align}} = \sum_{i<j} A_i A_j I_{ij}, \tag{5}$$

which prefers stylistic combinations that mutually reinforce each other and fit the contextual distribution.

Coherent Superposition. Computationally, this stage represents **multi-dimensional feature aggregation and decoding**. Using the optimized style parameters, the generator produces multiple candidate comments by sampling from the aggregated stylistic space. Each candidate is represented in the latent space as:

$$\mathbf{\Psi}_k = \mathbf{\Psi}_0 + \sum_{i=1}^{4} A^*_{k,i} \mathbf{d}^*_{k,i} + \epsilon_k, \tag{6}$$

where $A^*_{k,i}$ and $\mathbf{d}^*_{k,i}$ reflect the actual stylistic features realized during text decoding, and ϵ_k denotes minor sampling variations. To evaluate the coherence of each candidate with respect to community preference, we compute a standard cosine similarity score between the realized representation and the baseline preference:

$$C_k = \frac{\mathbf{\Psi}_k \cdot \mathbf{\Psi}_0}{\|\mathbf{\Psi}_k\| \, \|\mathbf{\Psi}_0\|}. \tag{7}$$

Higher scores indicate that the generated candidate naturally resonates with the historical community tendencies.

Collapse and Emission. Finally, acting as a **resonance-aware candidate selection** mechanism, this stage finalizes the text output. In probabilistic text generation, a model's continuous probability distribution effectively "collapses" into a single discrete sequence. We select the candidate with the highest coherence:

$$\mathbf{\Psi}^* = \arg\max_k C_k. \tag{8}$$

The selected comment undergoes minimal refinement for clarity and safety. The resulting output represents the expression that best aligns with the community's stylistic preferences, successfully materializing the computed social resonance into readable text.

6 Experiments

6.1 Evaluation Metrics

We evaluate model performance along three dimensions: **content quality**, **popularity prediction**, and **user behavior simulation**.

Content Quality. To assess semantic adequacy under the open-ended, multi-reference nature of comment generation, we compute similarity exclusively against **popular comments** from the dataset. We follow a weighted multi-reference strategy inspired by W-BLEU[59], where each reference comment is assigned a weight determined by its real-world engagement level. Engagement values are mapped onto a Gaussian distribution constrained to [0.6,1.0], giving more influential references a higher contribution while preserving diversity among less salient ones. The final score for each generated comment is the weighted maximum similarity across BLEU-1, METEOR, and BERTScore F1 For **stylistic evaluation**, we employ both *Qwen3-14B* (using a publicly released checkpoint) and *GPT-4o* as independent expert evaluators. Scores from the two models are averaged to reduce evaluator bias and avoid collapsing stylistic judgments onto the preference of any single evaluator. We report dimension-wise scores on Linguistic Expression, Creative Imagination, Emotional Resonance, and Social/Cultural Influence, and their mean forms the **Stylistic Resonance Score (SRS)**.

Models	Content Quality				Popularity	UBS
	BLEU-1	METEOR	F1	SRS	Prediction	
Mistral-7B	8.11	9.75	53.89	49.58	69.63	63.71
Baichuan2-7B	7.34	8.94	57.04	39.58	57.39	31.68
Qwen2.5-0.5B	14.22	13.41	57.79	31.34	66.69	62.23
Qwen2.5-7B	17.08	17.14	58.63	45.17	75.84	76.71
Qwen2.5-14B	15.36	15.64	58.02	51.78	76.70	70.19
+StyleCmt	21.08	18.61	63.18	60.78	82.18	84.50
LLaMA3.1-8B	17.03	16.64	57.78	47.66	47.25	63.99
+StyleCmt	20.00	20.48	60.90	57.74	73.52	71.49
ChatGPT-4o	16.97	16.41	59.78	57.09	71.41	72.38
+StyleCmt	20.57	21.15	63.98	59.49	76.98	77.69

Table 2: Results of text-based comment generation using large language models (LLMs). This table reports quantitative results of baseline and enhanced models on the HotComment benchmark. F1 denotes the BERTScore F1 metric. SRS is the averaged stylistic quality score derived from four stylistic dimensions (*LE, CI, ER, SCI*) under Content Quality. UBS represents the User Behavior Simulation score reflecting user-level response patterns. Cells highlighted in color indicate the top-three results within each column: 1st , 2nd , and 3rd .

Popularity Prediction. To measure alignment with real engagement patterns, semantic features from the generated comment are combined with **textual metadata** from the source content (title, keywords, and description). A trained prediction model outputs a normalized popularity score that reflects expected audience interaction. The predictor itself is reliable, achieving an accuracy of 82.61 and an F1 score of 79.39; details and full results are provided in the Appendix.

User Behavior Simulation. To approximate user reactions, we perform agent-based simulation using Qwen3-14B. Given a piece of content and a generated comment, the simulator estimates an **engagement score** reflecting the likelihood of user interaction, offering a behavioral view of comment effectiveness. As additional evidence of validity, on a within-item ranking test the simulator achieves a mean NDCG of 70.13 with Qwen3-14B and 68.34 with ChatGPT-4o as evaluators; full experimental details are deferred to the Appendix.

6.2 Experimental Setups

Dataset Partitioning. To ensure a reliable and unbiased evaluation, the dataset is divided into **training**, **validation**, and **test** sets following an 8:1:1 ratio. To prevent temporal and topical leakage, samples are stratified according to both **publication time** and

Models	Content Quality				Popularity	UBS
	BLEU-1	METEOR	F1	SRS	Prediction	
Mistral-3.1-24B	15.32	16.07	60.68	54.25	74.22	76.08
Gemini2.5-Image	15.25	13.69	61.34	56.63	72.05	73.09
ChatGPT-4o	14.84	12.65	61.21	55.71	72.06	73.01
ChatGPT-4o-mini	13.25	10.47	61.21	52.56	70.03	71.07
Qwen3-VL-4B	12.75	13.31	58.42	34.38	68.03	64.22
Qwen3-VL-8B	14.55	15.59	59.03	49.81	72.52	70.58
+StyleCmt	19.27	19.53	62.78	60.98	82.06	84.09
LLaMA3.2-Vision	13.82	13.84	60.04	49.43	62.03	65.08
+StyleCmt	17.58	16.89	62.57	57.42	75.07	75.04

Table 3: Results of multimodal comment generation using vision-language models (MLLMs). F1 denotes the BERTScore F1 metric. SRS is the averaged stylistic quality score derived from four stylistic dimensions (*LE, CI, ER, SCI*) under Content Quality. UBS represents the User Behavior Simulation score reflecting user-level response patterns. Cells highlighted in color indicate the top-three results within each column: 1st, 2nd, and 3rd.

Models	Content Quality				Popularity	UBS
	BLEU-1	METEOR	F1	SRS	Prediction	
Qwen2.5-14B	15.36	15.64	58.02	51.78	76.70	70.19
+CoT	17.01 (+10.74%)	16.64 (+6.39%)	59.02 (+1.72%)	54.35 (+4.96%)	78.95 (+2.93%)	79.82 (+13.72%)
+5-shot	17.80 (+15.89%)	16.27 (+4.03%)	60.92 (+5.00%)	55.72 (+7.61%)	77.37 (+0.87%)	76.77 (+9.37%)
+StyleCmt (Ours)	21.08 (+37.24%)	18.61 (+18.99%)	63.18 (+8.89%)	60.78 (+17.38%)	82.18 (+7.14%)	84.50 (+20.39%)
LLaMA3.1-8B	17.03	16.64	57.78	47.66	47.25	63.99
+CoT	17.09 (+0.35%)	16.53 (-0.66%)	60.38 (+4.50%)	47.83 (+0.36%)	59.79 (+26.54%)	67.99 (+6.25%)
+5-shot	18.55 (+8.93%)	17.58 (+5.65%)	60.31 (+4.38%)	52.89 (+10.97%)	67.54 (+42.94%)	70.33 (+9.91%)
+StyleCmt (Ours)	20.00 (+17.44%)	20.48 (+23.08%)	60.90 (+5.40%)	57.74 (+21.15%)	73.52 (+55.60%)	71.49 (+11.72%)

Table 4: Comparison of StyleCmt with reasoning-based methods on text-based comment generation (LLMs). Each enhancement row reports the absolute score (top) and relative improvement (bottom, in parentheses). SRS denotes the averaged stylistic resonance score, and UBS represents simulated user engagement. A continuous color gradient is applied within each model group to indicate the magnitude of performance gain, where deeper shading reflects greater relative improvement.

content category, ensuring that later-published content does not share topics with earlier training data. This temporal–categorical balance effectively mitigates overfitting caused by event recency and maintains domain diversity across splits.

Implementation Details. The *Popularity Prediction* model in our benchmark is trained on the training split, while evaluation is performed exclusively on the held-out test set. To ensure fairness, all open-source models are initialized from their official instruction-tuned checkpoints, and no additional fine-tuning or task adaptation is applied during evaluation. Experiments are conducted on **NVIDIA A100 GPUs**. All open-source models are deployed with **int8 quantization** to enable efficient inference within GPU memory constraints. For closed-source models such as GPT-4o, we access them through standardized API interfaces to maintain consistent prompt formatting and evaluation settings across modalities.

6.3 Benchmark Results

Text-based Comment Generation (LLMs). We evaluate a series of large language models, including *Qwen2.5*[75], *LLaMA3.1*[16], *ChatGPT-4o*[56], *Mistral*[29], and *Baichuan2*[74], on the text-based comment generation task. Tab. 2 shows that StyleCmt improves text-based comment generation across all evaluated large language models. Baseline models already demonstrate strong semantic quality, yet they often produce comments that lack stylistic alignment or social relevance. With StyleCmt applied, all LLMs exhibit consistent gains in both semantic similarity metrics and stylistic quality measures. Models such as Qwen2.5 and LLaMA3.1 show clear increases in BLEU-1, METEOR, F1, and SRS, while engagement-oriented metrics including Popularity Prediction and UBS also rise. The improvements are stable across model sizes and architectures, indicating that StyleCmt provides a generalizable enhancement to text-based comment generation.

Multimodal Comment Generation (MLLMs). We further evaluate multimodal large language models, including *Qwen3-VL*[75], *LLaMA3.2-Vision*[33], *Mistral-3.1*[29], *Gemini2.5-Image*[15], *ChatGPT-4o*, and *ChatGPT-4o-mini*[56], on video–comment generation under the same test configuration. A similar trend appears in video-based generation. As reported in Tab. 3, multimodal models benefit from StyleCmt with consistent improvements in semantic accuracy, stylistic coherence, and engagement-related scores. Qwen3-VL and LLaMA3.2-Vision both show sizeable gains across BLEU-1, METEOR, SRS, and UBS, demonstrating that stylistic conditioning enhances multimodal grounding as well. Notably, open-source models equipped with StyleCmt achieve performance comparable to or exceeding certain closed-source baselines in several engagement metrics, suggesting that audience-aware stylistic modeling can compensate for differences in model scale.

6.4 Performance of StyleCmt

Comprehensive Comparison across Models. Across both LLMs and MLLMs, StyleCmt provides consistent improvements over all evaluated baselines. The gains extend across semantic similarity, stylistic quality, and engagement-focused metrics. Larger instruction-tuned models tend to benefit more from StyleCmt, although smaller models also exhibit notable increases. The overall pattern indicates that

Models	Content Quality				Popularity	UBS
	BLEU-1	METEOR	F1	SRS	Prediction	
Qwen3-VL-8B	14.55	15.59	59.03	49.81	72.52	70.58
+CoT	15.46 (+6.25%)	16.17 (+3.72%)	60.93 (+3.22%)	52.88 (+6.16%)	78.27 (+7.93%)	75.54 (+7.03%)
+5-shot	16.52 (+13.54%)	16.81 (+7.83%)	61.59 (+4.34%)	54.79 (+10.00%)	78.87 (+8.76%)	76.25 (+8.03%)
+StyleCmt (Ours)	19.27 (+32.44%)	19.53 (+25.27%)	62.78 (+6.35%)	60.98 (+22.43%)	82.06 (+13.15%)	84.09 (+19.14%)
LLaMA3.2-Vision-11B	13.82	13.84	60.04	49.43	62.03	65.08
+CoT	14.68 (+6.22%)	14.29 (+3.25%)	61.83 (+2.98%)	50.98 (+3.14%)	65.04 (+4.85%)	70.07 (+7.67%)
+5-shot	15.62 (+13.02%)	14.87 (+7.44%)	61.89 (+3.08%)	54.44 (+10.14%)	70.01 (+12.86%)	72.53 (+11.45%)
+StyleCmt (Ours)	17.58 (+27.21%)	16.89 (+22.04%)	62.57 (+4.21%)	57.42 (+16.16%)	75.07 (+21.02%)	75.04 (+15.30%)

Table 5: Comparison of StyleCmt with reasoning-based methods on video-based comment generation (MLLMs). Each enhancement row reports the absolute score (top) and relative improvement (bottom, in parentheses). SRS denotes the averaged stylistic resonance score, and UBS represents simulated user engagement. A continuous color gradient is applied within each model group to indicate the magnitude of performance gain, where deeper shading reflects greater relative improvement.

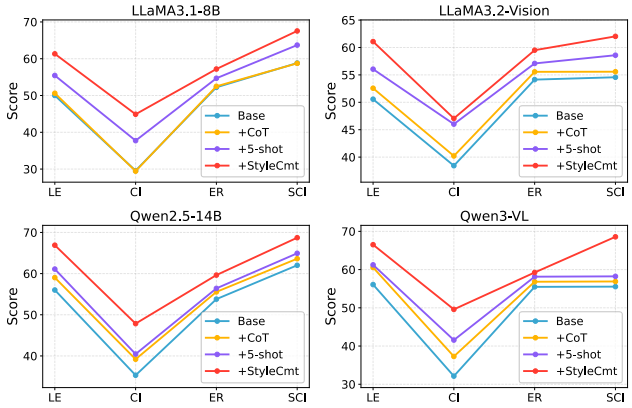

Figure 4: Improvements across stylistic dimensions. Relative gains of StyleCmt on four dimensions of the Stylistic Resonance Score (SRS): Linguistic Expression (LE), Creative Imagination (CI), Emotional Resonance (ER), and Social and Cultural Influence (SCI). The upward trend demonstrates balanced enhancement of stylistic coherence and expressiveness.

the proposed framework enhances linguistic expressiveness and social alignment in a model-agnostic manner, improving both content quality and predicted audience response.

Comparison with Other Strategies. We compare StyleCmt with two commonly used enhancement strategies: chain-of-thought prompting and few-shot prompting. Table4 and 5 show that while these strategies provide moderate gains in several metrics, their improvements are generally limited and inconsistent. Chain-of-thought tends to increase semantic coherence, and the five-shot setting offers slightly larger gains in some cases, yet both methods produce only small changes in stylistic resonance and user engagement, with increases in UBS remaining within a narrow range.

In contrast, StyleCmt consistently produces larger and more stable improvements across all evaluated models. It yields substantial gains in BLEU-1, METEOR, SRS, Popularity Prediction, and UBS, often exceeding the increases achieved by chain-of-thought or few-shot prompting by a wide margin. A similar pattern appears in LLaMA3.1 and Qwen3-VL, where StyleCmt provides stronger improvements across content quality and stylistic coherence. Moreover, StyleCmt enhances engagement-related metrics more effectively, indicating that the generated comments align more closely with expressive and affective tendencies commonly observed in real discussion environments.

These findings show that conventional enhancement strategies primarily improve local semantic refinement but are limited in their ability to capture stylistic preference patterns or broader discourse resonance. StyleCmt directly models such preferences and their interaction with social context, resulting in comments that are more expressive, contextually appropriate, and more effective at eliciting engagement. This demonstrates that stylistic conditioning plays a central role in improving the communicative impact of generated comments in social media settings.

Stylistic Resonance Analysis. To further examine the stylistic impact of StyleCmt, we analyze the relative improvements across the four dimensions that constitute the **Stylistic Resonance Score (SRS)**: **Linguistic Expression (LE)**, **Creative Imagination (CI)**, **Emotional Resonance (ER)**, and **Social and Cultural Influence (SCI)**. Figure 4 illustrates the percentage increases achieved by StyleCmt on each dimension for representative models. The results show steady gains across all stylistic aspects, confirming that StyleCmt enhances stylistic expressiveness in a balanced and interpretable manner.

The most substantial improvements are observed in *Creative Imagination* and *Social and Cultural Influence*, where StyleCmt strengthens associative creativity and contextual relevance to audience culture. *Linguistic Expression* and *Emotional Resonance* also exhibit clear upward trends, reflecting smoother rhetorical structure and more natural affective tone. The consistent growth across dimensions suggests that StyleCmt amplifies stylistic coherence rather than optimizing isolated features.

Overall, these results demonstrate that StyleCmt effectively models constructive interaction among stylistic components, leading to coordinated enhancement across expressive, imaginative, emotional, and social dimensions. The observed upward trajectories in the line chart highlight that stylistic resonance contributes directly to more engaging and audience-aligned comment generation.

7 Conclusion

In this work, we introduced **HotComment**, a large-scale benchmark for evaluating online comment generation across both textual and visual modalities. The benchmark integrates three complementary dimensions, including **Content Quality**, **Popularity Prediction**, and **User Behavior Simulation**, to assess linguistic expressiveness and social engagement potential in a unified framework. We further proposed **StyleCmt**, a wave-interference-inspired framework that models stylistic interactions to generate comments aligned with audience preferences. Extensive experiments demonstrate that StyleCmt consistently enhances content quality, stylistic richness, and engagement alignment across both LLMs and MLLMs, establishing a solid foundation for studying socially resonant comment generation.

References

[1] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada A. Adamic. 2012. The Role of Social Networks in Information Diffusion. In *Proceedings of the 21st International Conference on World Wide Web*. 519–528. doi:10.1145/2187836.2187907

[2] Jonah Berger and Katherine L. Milkman. 2012. What Makes Online Content Viral? *Journal of Marketing Research* 49, 2 (2012), 192–205. doi:10.1509/jmr.10.0353

[3] Qi Cao, Huawei Shen, Jinhua Gao, Bingzheng Wei, and Xueqi Cheng. 2020. Popularity Prediction on Social Platforms with Coupled Graph Neural Networks. In *Proceedings of the 13th ACM International Conference on Web Search and Data Mining*. 70–78.

[4] Yupeng Chang, Yi Chang, and Yuan Wu. 2026. BA-LoRA: Bias-Alleviating Low-Rank Adaptation to Mitigate Catastrophic Inheritance in Large Language Models. In *The Fourteenth International Conference on Learning Representations*. https://openreview.net/forum?id=q0X9SiXiRO

[5] Yupeng Chang, Yi Chang, and Yuan Wu. 2026. Decomposing Prompts: Discovering Reusable Scaffolds and Task-Specific Residuals. https://openreview.net/forum?id=WrTjCHs2tS

[6] Yupeng Chang, Chenlu Guo, Yi Chang, and Yuan Wu. 2025. LoRA-MGPO: Mitigating Double Descent in Low-Rank Adaptation via Momentum-Guided Perturbation Optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. 648–659.

[7] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology* 15, 3 (2024), 1–45.

[8] Jieting Chen, Junkai Ding, Wenping Chen, and Qin Jin. 2023. Knowledge Enhanced Model for Live Video Comment Generation. arXiv:2304.14657 [cs.CL]

[9] Yuyan Chen, Zhixu Li, Jiaqing Liang, Yanghua Xiao, Bang Liu, and Yunwen Chen. 2023. Can Pre-trained Language Models Understand Chinese Humor?. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining* (Singapore, Singapore) *(WSDM '23)*. Association for Computing Machinery, New York, NY, USA, 465–480. doi:10.1145/3539597.3570431

[10] Yuyan Chen, Songzhou Yan, Qingpei Guo, Jiyuan Jia, Zhixu Li, and Yanghua Xiao. 2024. HOTVCOM: Generating Buzzworthy Comments for Videos. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 2198–2224. doi:10.18653/v1/2024.findings-acl.130

[11] Yuyan Chen, Yichen Yuan, Panjun Liu, Dayiheng Liu, Qinghao Guan, Mengfei Guo, Haiming Peng, Bang Liu, Zhixu Li, and Yanghua Xiao. 2024. Talk Funny! A Large-Scale Humor Response Dataset with Chain-of-Humor Interpretation. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 16 (Mar. 2024), 17826–17834. doi:10.1609/aaai.v38i16.29736

[12] Zhiwei Chen, Yupeng Hu, Zhiheng Fu, Zixu Li, Jiale Huang, Qinlei Huang, and Yinwei Wei. 2026. Intent: Invariance and discrimination-aware noise mitigation for robust composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 20463–20471.

[13] Zhiwei Chen, Yupeng Hu, Zixu Li, Zhiheng Fu, Xuemeng Song, and Liqiang Nie. 2025. Offset: Segmentation-based focus shift revision for composed image retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 6113–6122.

[14] Justin Cheng, Lada A. Adamic, P. Alex Dow, Jon Kleinberg, and Jure Leskovec. 2014. Can Cascades Be Predicted?. In *Proceedings of the 23rd International Conference on World Wide Web*. 925–936. doi:10.1145/2566486.2567997

[15] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261* (2025).

[16] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI] https://arxiv.org/abs/2407.21783

[17] Emilio Ferrara and Zeyao Yang. 2015. Measuring Emotional Contagion in Social Media. *PLOS ONE* 10, 11 (11 2015), 1–14. doi:10.1371/journal.pone.0142390

[18] Zhiheng Fu, Yupeng Hu, Qianyun Yang, Shiqi Zhang, Zhiwei Chen, and Zixu Li. 2026. Air-Know: Arbiter-Calibrated Knowledge-Internalizing Robust Network for Composed Image Retrieval. arXiv:2604.19386 [cs.CV] https://arxiv.org/abs/2604.19386

[19] Soichiro Fujita, Hayato Kobayashi, and Manabu Okumura. 2019. Dataset Creation for Ranking Constructive News Comments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2619–2626. doi:10.18653/v1/P19-1250

[20] R. Kelly Garrett. 2009. Echo Chambers Online?: Politically Motivated Selective Exposure among Internet News Users. *Journal of Computer-Mediated Communication* 14, 2 (2009), 265–285. doi:10.1111/j.1083-6101.2009.01440.x

[21] R. Kelly Garrett, Dustin Carnahan, and Emily K. Lynch. 2013. A Turn Toward Avoidance? Selective Exposure to Online Political Information, 2004–2008. *Political Behavior* 35, 1 (2013), 113–134. doi:10.1007/s11109-011-9185-6

[22] Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan J. Watts. 2016. The Structural Virality of Online Diffusion. *Management Science* 62, 1 (2016), 180–196. doi:10.1287/mnsc.2015.2158

[23] Swapna Gottipati and Jing Jiang. 2012. Finding Thoughtful Comments from Social Media. In *Proceedings of COLING 2012: Technical Papers*. 995–1010. doi:10.3115/2380863.2380885

[24] Ruiqi He, Yushu He, Longju Bai, Jiarui Liu, Zhenjie Sun, Zenghao Tang, He Wang, Hanchen Xia, Rada Mihalcea, and Naihao Deng. 2025. Chumor 2.0: Towards Better Benchmarking Chinese Humor Understanding from (Ruo Zhi Ba). In *Findings of the Association for Computational Linguistics: ACL 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 21799–21818. doi:10.18653/v1/2025.findings-acl.1122

[25] Eugene Hecht. 2016. *Optics* (5th ed.). Pearson Education.

[26] Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. Do Androids Laugh at Electric Sheep? Humor "Understanding" Benchmarks from The New Yorker Caption Contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 688–714. doi:10.18653/v1/2023.acl-long.41

[27] Yupeng Hu, Zixu Li, Zhiwei Chen, Qinlei Huang, Zhiheng Fu, Mingzhu Xu, and Liqiang Nie. 2026. Refine: Composed video retrieval via shared and differential semantics enhancement. *ACM Transactions on Multimedia Computing, Communications and Applications* (2026).

[28] Yangliu Hu, Zikai Song, Na Feng, Yawei Luo, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. 2025. Sf2t: Self-supervised fragment finetuning of video-llms for fine-grained understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 29108–29117.

[29] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. arXiv:2310.06825 [cs.CL] https://arxiv.org/abs/2310.06825

[30] Elihu Katz, Jay G Blumler, and Michael Gurevitch. 1973. Uses and gratifications research. *The public opinion quarterly* 37, 4 (1973), 509–523.

[31] Hayato Kobayashi, Hiroaki Taguchi, Yoshimune Tabuchi, Chahine Koleejan, Ken Kobayashi, Soichiro Fujita, Kazuma Murao, Takeshi Masuyama, Taichi Yatsuka, Manabu Okumura, and Satoshi Sekine. 2021. A Case Study of In-House Competition for Ranking Constructive Comments in a News Service. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*. 24–35. doi:10.18653/v1/2021.socialnlp-1.3

[32] Varada Kolhatkar and Maite Taboada. 2017. Constructive Language in News Comments. In *Proceedings of the First Workshop on Abusive Language Online*. 11–17. doi:10.18653/v1/W17-3002

[33] Jewon Lee, Ki-Ung Song, Seungmin Yang, Donguk Lim, Jaeyeon Kim, Wooksu Shin, Bo-Kyeong Kim, Yong Jae Lee, and Tae-Ho Kim. 2025. Efficient LLaMA-3.2-Vision by Trimming Cross-attended Visual Features. arXiv:2504.00557 [cs.CV] https://arxiv.org/abs/2504.00557

[34] Jae Kook Lee and Eunyi Kim. 2017. Incidental Exposure to News: Predictors in the Social Media Setting and Effects on Information Gain Online. *Computers in Human Behavior* 75 (2017), 1008–1015. doi:10.1016/j.chb.2017.02.018

[35] Yiming Lei, Chenkai Zhang, Zeming Liu, Haitao Leng, ShaoGuo Liu, Tingting Gao, Qingjie Liu, and Yunhong Wang. 2025. GODBench: A Benchmark for Multimodal Large Language Models in Video Comment Art. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 11884–11952. doi:10.18653/v1/2025.acl-long.583

[36] Kun Li, Chenwei Dai, Wei Zhou, and Songlin Hu. 2024. Fine-grained User Behavior Simulation on Social Media Based on Role-playing Large Language Models. arXiv:2412.03148 [cs.CL]

[37] Wenjing Li, Zhongyuan Huang, Rui Xu, Ming Zhou, and Xiaolong Yang. 2019. Graph-to-Sequence Learning for News Comment Generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 455–465.

[38] Wenbing Li, Zikai Song, Jielei Zhang, Tianhao Zhao, Junkai Lin, Yiran Wang, and Wei Yang. 2026. Large Language Model as Token Compressor and Decompressor. arXiv:2603.25340 [cs.CL]

[39] Wenbing Li, Zikai Song, Hang Zhou, Yunyao Zhang, Junqing Yu, and Wei Yang. 2025. LoRA-Mixer: Coordinate Modular LoRA Experts Through Serial Attention Routing. *arXiv preprint arXiv:2507.00029* (2025).

[40] Wenbing Li, Hang Zhou, Junqing Yu, Zikai Song, and Wei Yang. 2024. Coupled mamba: Enhanced multimodal fusion with coupled state space model. *Advances in Neural Information Processing Systems* 37 (2024), 59808–59832.

[41] Yanshu Li, Yi Cao, Hongyang He, Qisen Cheng, Xiang Fu, Xi Xiao, Tianyang Wang, and Ruixiang Tang. 2025. M²IV: Towards Efficient and Fine-grained Multimodal In-Context Learning via Representation Engineering. In *Second Conference on Language Modeling*. https://openreview.net/forum?id=9ffYcEiNw9

[42] Yanshu Li, Jianjiang Yang, Tian Yun, Pinyuan Feng, Jinfa Huang, and Ruixiang Tang. 2025. Taco: Enhancing multimodal in-context learning via task mapping-guided sequence configuration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 736–763.

[43] Zixu Li, Zhiwei Chen, Haokun Wen, Zhiheng Fu, Yupeng Hu, and Weili Guan. 2025. Encoder: Entity mining and modification relation binding for composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 5101–5109.

[44] Zixu Li, Yupeng Hu, Zhiwei Chen, Qinlei Huang, Guozhi Qiu, Zhiheng Fu, and Meng Liu. 2026. Retrack: Evidence-driven dual-stream directional anchor calibration network for composed video retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 23373–23381.

[45] Zixu Li, Yupeng Hu, Zhiwei Chen, Mingyu Zhang, Zhiheng Fu, and Liqiang Nie. 2026. ConeSep: Cone-based Robust Noise-Unlearning Compositional Network for Composed Image Retrieval. arXiv:2604.20358 [cs.CV] https://arxiv.org/abs/2604.20358

[46] Zixu Li, Yupeng Hu, Zhiwei Chen, Shiqi Zhang, Qinlei Huang, Zhiheng Fu, and Yinwei Wei. 2026. Habit: Chrono-synergia robust progressive learning framework for composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 6762–6770.

[47] Zixu Li, Yupeng Hu, Zhiheng Fu, Zhiwei Chen, Yongqi Li, and Liqiang Nie. 2026. TEMA: Anchor the Image, Follow the Text for Multi-Modification Composed Image Retrieval. arXiv:2604.21806 [cs.CV] https://arxiv.org/abs/2604.21806

[48] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013/

[49] Xudong Lin, Ali Zare, Shiyuan Huang, Ming-Hsuan Yang, Shih-Fu Chang, and Li Zhang. 2024. Personalized Video Comment Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 16806–16820. doi:10.18653/v1/2024.findings-emnlp.979

[50] Yucheng Lin, Yuhan Xia, and Yunfei Long. 2024. Augmenting emotion features in irony detection with Large language modeling. arXiv:2404.12291 [cs.CL] https://arxiv.org/abs/2404.12291

[51] Ge Luo, Yuchen Ma, Manman Zhang, Junqiang Huang, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. 2024. Engaging Live Video Comments Generation. In *Proceedings of the 32nd ACM International Conference on Multimedia* (Melbourne VIC, Australia) *(MM '24)*. Association for Computing Machinery, New York, NY, USA, 8034–8042. doi:10.1145/3664647.3681195

[52] Shuming Ma, Lei Cui, Damai Dai, Furu Wei, and Xu Sun. 2019. LiveBot: Generating Live Video Comments Based on Visual and Textual Contexts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 6810–6817. doi:10.1609/aaai.v33i01.33016810

[53] Xuanjing Ma, Xiaochun Li, Lei Wang, and Wei Zhang. 2018. Unsupervised Article Comment Generation with Semantic Matching. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1234–1243.

[54] Zixiang Meng, Qiang Gao, Di Guo, Yunlong Li, Bobo Li, Hao Fei, Shengqiong Wu, Fei Li, Chong Teng, and Donghong Ji. 2024. MMLSCU: A Dataset for Multi-modal Multi-domain Live Streaming Comment Understanding. In *Proceedings of the ACM on Web Conference 2024*. 4395–4406. doi:10.1145/3589334.3645677

[55] Matteo Migliarini, Berat Ercevik, Oluwagbemike Olowe, Saira Fatima, Sarah Zhao, Minh Anh Le, Vasu Sharma, and Ashwinee Panda. 2026. @grokSet: Multi-party Human-LLM Interactions in Social Media. arXiv:2602.21236 [cs.CL]

[56] OpenAI. 2023. GPT-4 Technical Report. In *arXiv preprint arXiv:2303.08774*.

[57] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Pierre Isabelle, Eugene Charniak, and Dekang Lin (Eds.). Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318. doi:10.3115/1073083.1073135

[58] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–22. doi:10.1145/3586183.3606763

[59] Lianhui Qin, Lemao Liu, Victoria Bi, Yan Wang, Xiaojiang Liu, Zhiting Hu, Hai Zhao, and Shuming Shi. 2018. Automatic Article Commenting: The Task and Dataset. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 899–905.

[60] Zhongyi Qiu, Hanjia Lyu, Wei Xiong, and Jiebo Luo. 2025. Can LLMs Simulate Social Media Engagement? A Study on Action-Guided Response Generation. arXiv:2502.12073 [cs.CL] doi:10.48550/arXiv.2502.12073

[61] Julian Risch and Ralf Krestel. 2020. Top Comment or Flop Comment? Predicting and Explaining User Engagement in Online News Discussions. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 579–589. doi:10.1609/icwsm.v14i1.7325

[62] Svenja Schäfer. 2023. Incidental News Exposure in a Digital Media Environment: A Scoping Review of Recent Research. *Annals of the International Communication Association* 47, 2 (2023), 242–260. doi:10.1080/23808985.2023.2169953

[63] Zikai Song, Run Luo, Lintao Ma, Ying Tang, Yi-Ping Phoebe Chen, Junqing Yu, and Wei Yang. 2025. Temporal coherent object flow for multi-object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 6978–6986.

[64] Zikai Song, Run Luo, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. 2023. Compact transformer tracker with correlative masked modeling. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 2321–2329.

[65] Zikai Song, Ying Tang, Run Luo, Lintao Ma, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. 2024. Autogenic language embedding for coherent point tracking. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 2021–2030.

[66] Zikai Song, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. 2022. Transformer tracking with cyclic shifting window attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8791–8800.

[67] Zikai Song, Junqing Yu, Yi-Ping Phoebe Chen, Wei Yang, and Xinchao Wang. 2026. Hypergraph-State Collaborative Reasoning for Multi-Object Tracking. arXiv:2604.12665 [cs.CV]

[68] Jiao Sun, Anjali Narayan-Chen, Shereen Oraby, Alessandra Cervone, Tagyoung Chung, Jing Huang, Yang Liu, and Nanyun Peng. 2022. ExPUNations: Augmenting Puns with Keywords and Explanations. arXiv:2210.13513 [cs.CL] https://arxiv.org/abs/2210.13513

[69] Yuchong Sun, Bei Liu, Xu Chen, Ruihua Song, and Jianlong Fu. 2024. ViCo: Engaging Video Comment Generation with Human Preference Rewards. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia (MMAsia '24)*. Association for Computing Machinery, New York, NY, USA, Article 98, 1 pages. doi:10.1145/3696409.3700260

[70] Petter Törnberg, Diliara Valeeva, Justus Uitermark, and Christopher Bail. 2023. Simulating Social Media Using Large Language Models to Evaluate Alternative News Feed Algorithms. arXiv:2310.05984 [cs.CY]

[71] Lilian Weng, Alessandro Flammini, Alessandro Vespignani, and Filippo Menczer. 2012. Competition among Memes in a World with Limited Attention. *Scientific Reports* 2 (2012), 335. doi:10.1038/srep00335

[72] Yijie Xu, Bolun Zheng, Wei Zhu, Hangjia Pan, Yuchen Yao, Ning Xu, Anan Liu, Quan Zhang, and Chenggang Yan. 2025. SMTPD: A New Benchmark for Temporal Prediction of Social Media Popularity. arXiv:2503.04446 [cs.SI]

[73] Zhijun Xu, Siyu Yuan, Lingjie Chen, and Deqing Yang. 2024. "A good pun is its own reword": Can Large Language Models Understand Puns?. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 11766–11782. doi:10.18653/v1/2024.emnlp-main.657

[74] Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, JunTao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2025. Baichuan 2: Open Large-scale Language Models. arXiv:2309.10305 [cs.CL] https://arxiv.org/abs/2309.10305

[75] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 Technical Report. arXiv:2412.15115 [cs.CL] https://arxiv.org/abs/2412.15115

[76] Fuchun Yang, Peng Li, Xin Guo, Yujie Zhang, and Shiyue Chen. 2019. Read-Attend-Comment: A Reading Comprehension Based Approach for Comment Generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2567–2577.

[77] Fuchun Yang, Peng Li, Qiang Wang, Yuan Xu, and Zhenhua Wei. 2019. Cross-Modal Comment Generation with Image and Text Features. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 3456–3465.

[78] Qianyun Yang, Zhiwei Chen, Yupeng Hu, Zixu Li, Zhiheng Fu, and Liqiang Nie. 2026. Stable: Efficient Hybrid Nearest Neighbor Search via Magnitude-Uniformity and Cardinality-Robustness. *arXiv preprint arXiv:2604.01617* (2026).

[79] Liliang Ye, Yunyao Zhang, Yafeng Wu, Yi-Ping Phoebe Chen, Junqing Yu, Wei Yang, and Zikai Song. 2025. MVP: Winning Solution to SMP Challenge 2025 Video Track. *arXiv preprint arXiv:2507.00950* (2025).

[80] Erxin Yu, Jing Li, and Chunpu Xu. 2024. PopALM: Popularity-Aligned Language Models for Social Media Trendy Response Prediction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 12867–12878.

[81] Wenhuan Zeng, Abulikemu Abuduweili, Lei Li, and Pengcheng Yang. 2019. Automatic Generation of Personalized Comment Based on User Profile. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. 229–235. doi:10.18653/v1/P19-2032

[82] Zehua Zeng, Neng Gao, Cong Xue, and Chenyang Tu. 2021. PLVCG: A Pretraining Based Model for Live Video Comment Generation. In *Advances in Knowledge Discovery and Data Mining – 25th Pacific-Asia Conference, PAKDD 2021, Proceedings, Part II*. 690–702. doi:10.1007/978-3-030-75765-6_55

[83] Jinghao Zhang, Yuting Liu, Wenjie Wang, Qiang Liu, Shu Wu, Liang Wang, and Tat-Seng Chua. 2025. Personalized Text Generation with Contrastive Activation Steering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 7128–7141.

[84] Mingyu Zhang, Zixu Li, Zhiwei Chen, Zhiheng Fu, Xiaowei Zhu, Jiajia Nie, Yinwei Wei, and Yupeng Hu. 2026. HINT: Composed Image Retrieval with Dual-Path Compositional Contextualized Network. *arXiv preprint arXiv:2603.26341* (2026).

[85] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. arXiv:1904.09675 [cs.CL] https://arxiv.org/abs/1904.09675

[86] Xinglang Zhang, Yunyao Zhang, ZeLiang Chen, Junqing Yu, Wei Yang, and Zikai Song. 2026. Logical Phase Transitions: Understanding Collapse in LLM Logical Reasoning. arXiv:2601.02902 [cs.AI]

[87] Yunyao Zhang, Yihao Ai, Zuocheng Ying, Qirui Mi, Junqing Yu, Wei Yang, and Zikai Song. 2026. Coupling Macro Dynamics and Micro States for Long-Horizon Social Simulation. arXiv:2604.05516 [cs.SI]

[88] Yongmao Zhang, Kai Qiao, Zhengyan Wang, Ningning Liang, Dekui Ma, Wenyao Sun, Jian Chen, and Bin Yan. 2026. POSIM: A Multi-Agent Simulation Framework for Social Media Public Opinion Evolution and Governance. arXiv:2603.23884 [cs.CL]

[89] Yunyao Zhang, Zikai Song, Hang Zhou, Wenfeng Ren, Yi-Ping Phoebe Chen, Junqing Yu, and Wei Yang. 2025. GA-S3: Comprehensive Social Network Simulation with Group Agents. In *Findings of the Association for Computational Linguistics: ACL 2025*. 8950–8970.

[90] Yunyao Zhang, Zuocheng Ying, Xinglang Zhang, Junqing Yu, Peng Fang, Xu Chen, Wei Yang, and Zikai Song. 2026. IntervenSim: Intervention-Aware Social Network Simulation for Opinion Dynamics. arXiv:2604.06600 [cs.SI]

[91] Yunyao Zhang, Xinglang Zhang, Junxi Sheng, Wenbing Li, Junqing Yu, Yi-Ping Phoebe Chen, Wei Yang, and Zikai Song. 2026. Semantic-Aware Logical Reasoning via a Semiotic Framework. arXiv:2509.24765 [cs.AI]

[92] Haitao Zheng, Wei Wang, Wang Chen, and Arun Kumar Sangaiah. 2018. Automatic Generation of News Comments Based on Gated Attention Neural Networks. In *IEEE Access*, Vol. 6. IEEE, 702–710.

[93] Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. 2024. Let's Think Outside the Box: Exploring Leap-of-Thought in Large Language Models with Creative Humor Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13246–13257.