Community-Aware Assessment of Social Textual Engagement and Resonance: A Human-Centric Perspective on User-Generated Content Evaluation

Anonymous ACL submission

Abstract

Traditional Video Quality Assessment (VQA) focuses narrowly on aesthetic fidelity, overlooking the complex social dynamics that define quality in User-Generated Content (UGC). In this work, we propose a paradigm shift from signal-centric metrics to human-centric resonance assessment. We introduce CASTER (Community-Aware Assessment of Social Textual Engagement and Resonance), a new task that evaluates whether a UGC item achieves positive community resonance based on its multimodal attributes rather than visual quality alone. To address this, we present MEDEA (Multimodal Engagement-Driven Evaluation Architecture), which introduces a novel Social Chain-of-Thought (Social-CoT) mechanism. Unlike traditional logical CoT, Social-CoT performs multimodal perspective-taking, instantiating diverse viewer personas to simulate collective cognitive and emotional reactions (i.e., the "community mind") before deriving a quality judgment. MEDEA is trained via a two-stage approach involving supervised fine-tuning and process-supervised reinforcement learning with Social Alignment Reward to ensure reasoning paths are grounded in authentic human social cognition. To support this task, we release CASTER-Bench, a comprehensive human-annotated benchmark covering diverse UGC categories. Experiments demonstrate that MEDEA significantly outperforms state-of-the-art baselines on CASTER-Bench while providing interpretable and empathetic reasoning paths that align with real community feedback.

1 Introduction

Traditional Video Quality Assessment (VQA) has achieved notable success in measuring aesthetic fidelity and technical distortions (Seshadrinathan et al., 2010; Lin et al., 2015; Danier et al., 2023). However, its core objective is fundamentally misaligned with how quality is perceived on User-Generated Content (UGC) platforms. By focusing primarily on pixel level integrity and low-level visual cues, existing VQA methods (Wu et al., 2023a,b; Lu et al., 2024; Duan et al., 2025) fail to capture the human-centered and social nature of quality in real-world UGC. As a result, these approaches struggle to reflect whether content is meaningful, engaging, or valuable to actual users beyond momentary visual appeal. [1]

The key challenge, therefore, lies in how to properly define UGC quality. On large-scale platforms, high-quality content is determined not by technical perfection, but by whether it resonates with the community eliciting emotional engagement, meaningful discussion, and positive recognition. Such community endorsement is most explicitly reflected through user engagement signals, among which positive comments provide direct, content level evidence of perceived quality.

While Large Language Models (LLMs) have demonstrated impressive reasoning capabilities via Chain-of-Thought (CoT) in logical and mathematical domains (Wei et al., 2022), *Social Reasoning*, the ability to model human emotional dynamics and collective reception remains underexplored. We argue that assessing UGC quality requires a Theory of Mind (ToM) approach (Sap et al., 2022): the model must not merely analyze the content signals, but actively "step into the shoes" of the audience. We term this process *Social Chain-of-Thought (Social-CoT)*, where the model explicitly generates a diverse set of empathetic reaction paths simulating the "community mind" before converging on a quality judgment.

Motivated by this observation, we introduce *CASTER* (Community-Aware Assessment of Social Textual Engagement and Resonance), a task that reframes UGC quality assessment as identifying content genuinely endorsed by its audience via

[1] Detailed related works can be found in Appendix A.

social reasoning.

However, direct access to user comments is often unavailable, especially for newly uploaded or sparsely interacted content, where quality assessment is still critically needed for recommendation and moderation. To address this limitation, we propose *MEDEA* (Multimodal Engagement-Driven Evaluation Architecture), which operationalizes the Social-CoT paradigm. MEDEA infers community resonance by instantiating diverse viewer personas and simulating plausible user comments conditioned on multimodal content signals, effectively performing multimodal perspective-taking before aggregating these reaction paths into a final quality judgment.

To achieve this capability, MEDEA is trained via supervised fine-tuning (SFT) and process-supervised reinforcement learning (RL), combining large-scale pseudo-labeled data with expert annotations. Crucially, we introduce *Social Alignment Reward* during the RL stage to ensure the generated reasoning paths are grounded in authentic human social cognition rather than robotic analysis. Experiments demonstrate that MEDEA substantially outperforms aesthetic and multimodal baselines (Wu et al., 2022, 2023b, 2024; Duan et al., 2025; Jia et al., 2025), while providing interpretable and community-aligned reasoning traces.

Furthermore, to support this task, we present *CASTER-Bench*, a multimodal benchmark specifically designed for long-form UGC videos, with an average duration of 442 seconds. Unlike existing VQA datasets that predominantly rely on short clips (typically 8-10 seconds), CASTER-Bench enables the evaluation of narrative coherence, information density, and sustained engagement that are critical in real-world content recommendation scenarios. The benchmark is annotated by expert raters using a human-centered rubric, and empirical analysis reveals a strong colorrelation between positive user comments and expert judgments, while traditional VQA and vision-centric models perform poorly. These results highlight the limitations of existing methods in modeling the semantic, social, and temporal factors underlying UGC quality.

Our contributions are summarized as follows:

- We introduce CASTER, a community-aware task that redefines UGC quality through the lens of social reasoning, and release CASTER-Bench, a multimodal benchmark annotated using a human-centered rubric.
- We propose MEDEA, an evaluation framework that pioneers Social-CoT to simulate empathetic user reactions, trained via SFT and process-supervised RL with Social Alignment Reward.
- We demonstrate that MEDEA significantly outperforms diverse types of baselines while offering improved interpretability through generated social reasoning paths.

2 Community-Aware Assessment of Social Textual Engagement and Resonance

This section formalizes the CASTER task and introduces CASTER-Bench, a benchmark designed to support this task. We describe the UGC item collection process, expert-driven annotation protocol, and quality control procedures, followed by dataset statistics and comparisons with existing benchmarks.

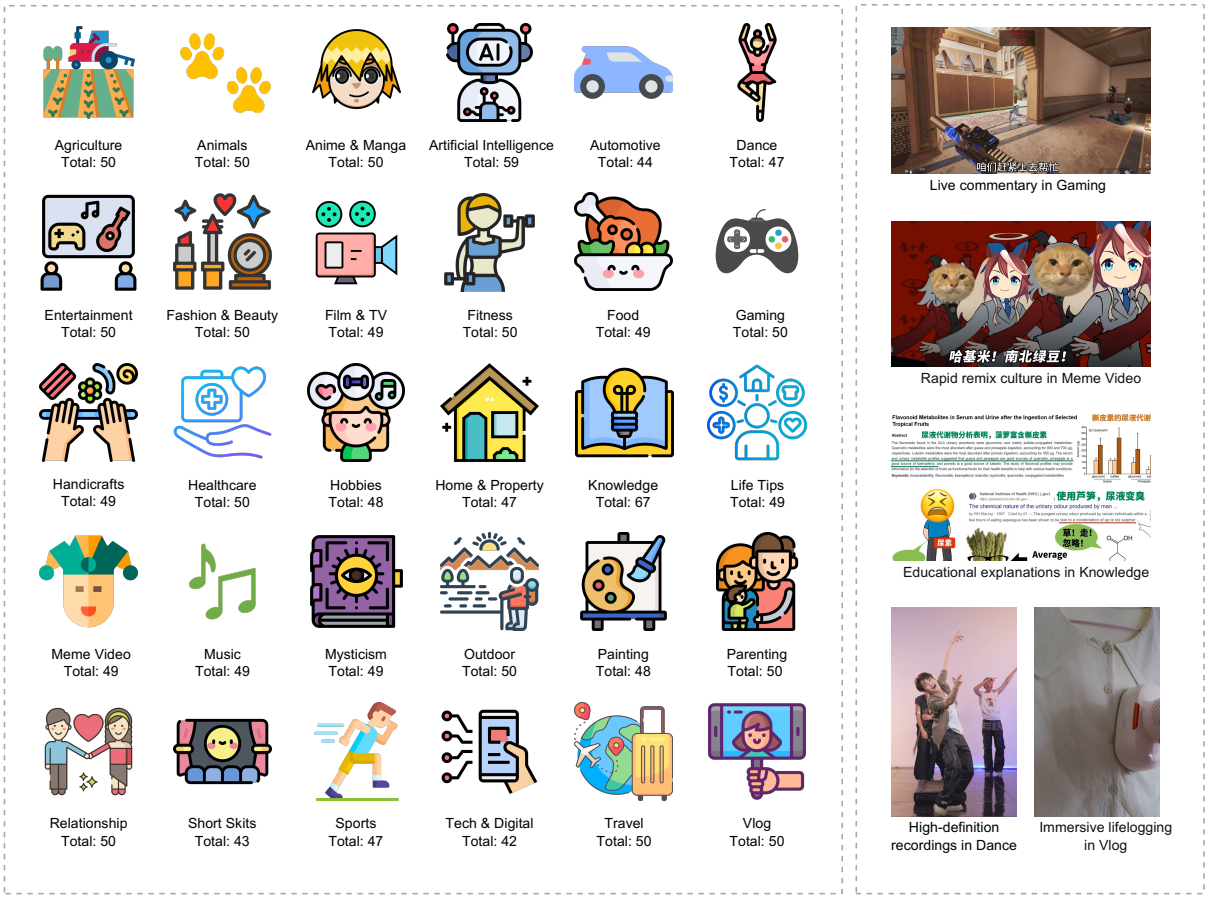
2.1 The CASTER Task

CASTER aims to assess whether a piece of user-generated content resonates with the community from a holistic, human-centric perspective. Unlike traditional video quality assessment which focuses on low-level aesthetic or technical attributes (e.g., sharpness or noise), CASTER evaluates the quality of the content artifact itself rather than the video signal alone.

Formally, given a UGC item consisting of multimodal inputs including video frames, cover image, title, tags, category metadata, and automatic speech recognition (ASR) transcripts, the task is to predict whether the content is perceived as *high-quality* or *low-quality* according to human judgment. This judgment reflects community-level resonance and is shaped by factors such as creativity, emotional engagement, informational value, narrative coherence, and originality. By framing quality assessment as a community-aware and content-driven task, CASTER decouples perceived quality from confounding signals such as view count or recommendation exposure, better aligning automatic evaluation with real user preferences.

2.2 CASTER-Bench: A Benchmark for Social Resonance

To support the CASTER task, we introduce CASTER-Bench, a human-annotated benchmark containing 1,485 UGC items curated from a large-scale comprehensive video platform and annotated by professional content operation experts.



(a) Category-level distribution of CASTER-Bench across 30 major UG categories. (b) Representative UGC examples.

Figure 1: Overview of CASTER-Bench. (a) Category-level composition of the benchmark, covering 1,485 UGC items sampled from 30 major content categories with balanced representation. (b) Representative examples illustrating diverse presentation styles and content paradigms, including live commentary, meme remix culture, educational explanations, high-definition performance recordings, and immersive vlogging.

In contrast to existing benchmarks such as KVQ (Lu et al., 2024) and FineVD (Duan et al., 2025), which emphasize aesthetic quality on short clips, CASTER-Bench focuses on subjective, multi-dimensional perceptions of long-form content quality (average 442s), including creativity, emotional value, informational utility, and narrative excellence. Each item is accompanied by rich multi-modal information, including visual content, cover image, title, tags, category metadata, and ASR transcripts, enabling holistic assessment beyond visual appearance alone.

2.2.1 Data Collection and Statistics

UGC items were collected following stratified random sampling across 30 major content categories (e.g., *Lifestyle*, *Knowledge*, *Gaming*) to ensure broad coverage of diverse content scenarios, as illustrated in Figure 1a. Figure 1b also demonstrates representative examples, highlighting the diversity in content forms and production paradigms.

CASTER-Bench contains 1,485 UGC items with a quality label distribution mirroring real-world platforms: *Excellent* (10.6%), *Good* (17.0%), *Average* (38.6%), and *Poor* (33.7%). This distribution presents a realistic challenge for identifying high-quality content amidst massive amounts of average data. Detailed statistics regarding technical attributes, such as duration and resolution, are provided in Appendix B.

2.2.2 Expert-Driven Annotation Protocol

To ensure the reliability, consistency, and practical relevance of the annotations, we adopt a rigorously designed expert-driven annotation protocol grounded in real-world content moderation and recommendation practices. In particular, we recruited 10 professional content operation experts to annotate the dataset. The annotation is based on a comprehensive framework comprising four core

Database	Source	Num.	Avg Dur.	Total Dur.	Focus	Modality	Env.
KoNViD-1k (Hosu et al., 2017)	YFCC100m	1200	8s	2.7h	A&T	Video Only	Crowd
LIVE-VQC (Sinno and Bovik, 2019)	UGC-P	585	10s	1.6h	A&T	Video Only	Crowd
YouTube-UGC (Wang et al., 2019)	UGC-P	1380	20s	7.7h	A&T	Video Only	Crowd
KVQ (Lu et al., 2024)	UGC-P	600	8s	8h	A&T	Video Only	In-lab
FineVD (Duan et al., 2025)	UGC-P	6104	8s	13.6h	A&T	Video Only	In-lab
CASTER-Bench (Ours)	UGC-P	1485	**442s**	**182.5h**	**S&C**	**T&T&V&A**	In-lab

Table 1: Multi-dimension comparison between mainstream general video quality assessment datasets. Num. denotes the total number of test video sequences; Avg Dur. and Total Dur. denote average duration per video (seconds) and combined duration of all videos (hours). A&T and S&C indicate aesthetic–technical and subjective content-driven quality; T&T&V&A includes title, tags, video, and ASR transcripts; Crowd and In-lab denote annotation environments.

dimensions:

- **Production Quality**: audiovisual execution, post-production, and special effects.
- **Perceived Value**: emotional resonance, entertainment, or affective engagement.
- **Information Utility**: practical knowledge, instructional value, or curated information.
- **Narrative Excellence**: coherent structure, originality, or innovative presentation.

Annotators labeled items as *Excellent*, *Good*, *Average*, or *Poor*. Crucially, they received high-engagement user comments and were instructed to use them as complementary evidence to judge whether content elicited genuine community resonance, rather than relying solely on visual signals.

A core objective of CASTER-Bench is to assess the *intrinsic value* of UGC rather than merely predicting popularity metrics like view counts, which are often saturated with noise such as recommendation biases and sensationalist tactics. The expert annotations serve as a "refinement" mechanism, filtering out confounding factors to prioritize genuine community resonance over superficial traffic. Detailed case studies distinguishing high-popularity content from high-quality content are provided in Appendix H. A sanitized version of the data will be provided in the final camera-ready version.

3 Multimodal Engagement-Driven Evaluation Architecture

In this section, we propose MEDEA, a unified framework that operationalizes the Social-CoT paradigm. Rather than mapping multimodal signals directly to a quality label, MEDEA simulates a "community of critics" by generating diverse empathetic reasoning paths before aggregating them into a final judgment. MEDEA follows a three-stage pipeline: (1) constructing a large-scale Social-CoT corpus by mining community reactions and instantiating viewer personas; (2) supervised fine-tuning to internalize the capability of multimodal perspective-taking; and (3) process-supervised reinforcement learning with Social Alignment Reward to refine the authenticity and diversity of the social reasoning process. Figure 2 provides an overview of the MEDEA framework.

3.1 Constructing Social-CoT Paths

To train a model capable of social reasoning, we construct a dataset that transforms raw UGC engagement signals into structured empathetic reasoning paths. We combine large-scale unlabeled scripts containing real user comments with a smaller, expert-annotated dataset.

Mining Community Reactions and Perspective Taking. We posit that understanding UGC quality requires identifying specific "viewer personas" within the community. Given a UGC item, we treat its comment section as a reflection of the collective "community mind". For unlabeled data, we retrieve the top-50 most-liked comments and employ a teacher model to filter for relevance, selecting 15-20 reactions that capture core dimensions such as creativity, emotional appeal, and narrative structure. These selected comments serve as authentic reaction anchors. For the reasoning process, we instruct Gemini-2.5-Flash to perform multimodal perspective-taking: it must instantiate diverse viewer personas and articulate why specific visual or narrative elements trigger specific reactions (refer to Appendix F for the detailed prompts). For data with expert-provided labels, we apply the same prompting pipeline but explicitly instruct the teacher model to ensure that both its reasoning process and final answer agree with the gold label.

Consensus Mechanism via Skellam Scoring. To transit from diverse social perspectives to a uni-

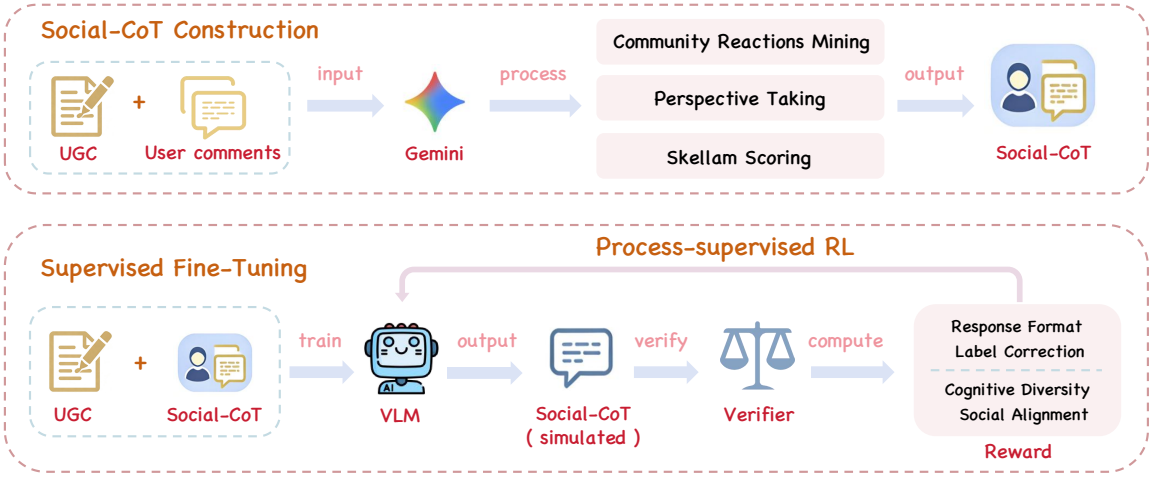


Figure 2: Overview of the MEDEA framework. The upper part depicts the Social-CoT construction pipeline, including community reactions mining, perspective taking, and consensus mechanism via Skellam Scoring. The lower part illustrates the training procedure, consisting of supervised fine-tuning and process-supervised reinforcement learning with multiple reward signals.

fied quality judgment, we implement a statistical consensus mechanism. Each reasoning path (simulated comment) is assigned a supportive or oppositional stance. Let X denote the number of supportive paths and Y denote the number of oppositional paths. We compute the Skellam-normalized difference score z to model the significance of the community endorsement:

$$z = \frac{X - Y}{\sqrt{X + Y}}. \tag{1}$$

A heuristic quality label is then assigned based on this community consensus:

$$\text{label} = \begin{cases} \text{High-Quality}, & \text{if } z \geq 1.5, \\ \text{Low-Quality}, & \text{otherwise.} \end{cases} \tag{2}$$

This "Think-then-Aggregate" structure forms the training target for our Social-CoT, ensuring the final judgment is causally derived from the simulated community dynamics.

3.2 Supervised Fine-Tuning for Social Reasoning

The first training stage involves Supervised Fine-Tuning (SFT) to teach the model the syntax and semantics of Social-CoT. We combine the heuristic-labeled Social-CoT data (from unlabeled UGC items) with human-annotated data into a unified corpus. SFT plays a crucial role in enabling multimodal grounding: it trains the model to align visual cues (e.g., lighting, editing pace) and textual metadata (titles, tags) with social interpretations. By learning to generate the reaction paths before predicting the label, the model internalizes a structured reasoning process, moving beyond black-box classification to interpretable social simulation.

3.3 Process-Supervised Reinforcement Learning

To further refine the quality of the Social-CoT generation, we employ Reinforcement Learning (RL) using Group Relative Policy Optimization (GRPO) (Shao et al., 2024). While SFT teaches the model how to reason, RL aligns the reasoning process with authentic human social cognition. We design a composite reward signal comprising four distinct components:

$$r = r_{\text{format}} + r_{\text{label}} + r_{\text{diversity}} + r_{\text{social}}. \tag{3}$$

Format and Label Rewards. r_{format} ensures the output adheres to the structured `<think>...</think>` format, while r_{label} rewards the correctness of the final binary quality prediction against the ground truth.

Cognitive Diversity Constraint ($r_{\text{diversity}}$). A robust community simulation should reflect a spectrum of opinions rather than repeating a single viewpoint. To prevent mode collapse where the model generates repetitive comments, we introduce a diversity penalty:

$$r_{\text{diversity}} = -\lambda_{\text{div}} \sum_{c \in \mathcal{C}} [f(c) - 1], \tag{4}$$

where $\mathcal{C}$ is the set of generated reaction paths and $f(c)$ denotes the frequency of identical or near-identical sentiments, forcing the model to explore the full distribution of potential audience reactions.

Social Alignment Reward (r_{social}). To ensure the generated reasoning paths are not hallucinations but are grounded in genuine human emotional expression, we introduce the Social Alignment Reward, which measures the semantic similarity between the model's simulated personas and real, high-engagement user comments from a held-out set. Let $\mathcal{G} = \{g_i\}$ be the set of generated reaction paths and $\mathcal{R} = \{r_j\}$ be the set of real user comments, we compute the cosine similarity between their embeddings:

$$S_{ij} = e(g_i)^\top e(r_j), \quad \text{where } e(x) = \frac{f(x)}{\|f(x)\|_2}. \tag{5}$$

We perform greedy matching to align each generated persona with the closest real-world counterpart. The final reward is the mean of these matched similarities:

$$r_{\text{social}} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} s. \tag{6}$$

This reward acts as a "social grounding" signal, encouraging the model to mimic the tone, nuance, and emotional granularity of actual community.

By combining all these rewards, the diversity and authenticity rewards ensure that simulated comments remain varied and semantically aligned with real user feedback, while the format and label rewards guarantee well-formed outputs and accurate final decisions. Together, these signals guide the model toward producing interpretable, community-grounded predictions for the CASTER task.

4 Experiments

In this section, we evaluate MEDEA on large-scale real-world UGC item assessment scenarios. We first introduce the experimental setups, including baselines and training data construction, followed by the main results on CASTER-Bench, and finally provide ablation studies to quantify the contribution of each system component.

4.1 Experimental Setups

Baselines. To comprehensively assess the performance of MEDEA, we compare it against a diverse set of baselines categorized into four groups:

1. **Traditional Video Quality Assessment (VQA) Methods:** This group includes representative regression-based models that focus on aesthetic and technical quality, including FastVQA (Wu et al., 2022), DOVER (Wu et al., 2023a), MaxVQA (Wu et al., 2023b), Q-Align (Wu et al., 2024), FineVQ (Duan et al., 2025), and VQA2 (Jia et al., 2025).
2. **Standard Large Multimodal Models (LMMs):** We evaluate general-purpose flagship models, including Qwen3-VL-Plus, GPT-5.2 and Claude-4.5-opus. These LMMs are the only flagship candidates capable of explicitly disabling the reasoning process, allowing us to establish a pure baseline for standard multimodal capabilities without intrinsic CoT interference.
3. **Reasoning-Enhanced LMMs (Long-CoT):** To benchmark against state-of-the-art intrinsic reasoning capabilities, we include models utilizing CoT or long-context reasoning. This category includes Qwen3-VL-8B-Think (the backbone of MEDEA), Qwen3-VL-Plus (reasoning), GPT-5.2 (reasoning), Gemini-3.0-Pro (reasoning), and Claude-4.5-opus (reasoning). For these models, we explicitly configured the reasoning effort to "high" to fully activate their extended thinking capabilities and maximize the depth of logical deduction.
4. **Flagship Models with Social-CoT Simulation:** To isolate the effectiveness of our proposed mechanism, we prompt non-reasoning models (Gemini-2.5-Flash, Qwen3-VL-Plus, and GPT-5.2) with the Social-CoT prompts used in MEDEA, forcing them to simulate social perspective-taking without fine-tuning.

For the Traditional VQA methods, which output continuous quality scores, we perform an exhaustive threshold sweep to map scores to binary classifications and report the best performance on CASTER-Bench, ensuring they are evaluated at their optimal operating points. Detailed results of these baselines across various thresholds are provided in Appendix I. All LMM-based baselines perform zero-shot prediction. Flagship Models with Social-CoT Simulation utilize the exact inference prompt as MEDEA to ensure a fair comparison of the reasoning framework itself. All reported results are averaged over five independent runs.

Training Data. The full data construction pipeline is described in Section 3.1. Here we

Method	High-Quality			Low-Quality			Macro Average		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Traditional VQA Methods									
FastVQA (Wu et al., 2022)	0.347	0.440	0.388	0.761	0.682	0.719	0.554	0.561	0.554
DOVER (Wu et al., 2023a)	0.308	0.377	0.339	0.739	0.676	0.706	0.524	0.526	0.523
MaxVQA (Wu et al., 2023b)	0.345	0.518	0.414	0.772	0.623	0.690	0.558	0.571	0.552
Q-Align (Wu et al., 2024)	0.382	0.404	0.392	0.766	0.749	0.758	0.574	0.577	0.575
FineVQ (Duan et al., 2025)	0.323	0.343	0.333	0.742	0.724	0.733	0.532	0.534	0.533
VQA2 (Jia et al., 2025)	0.358	0.454	0.400	0.766	0.688	0.725	0.562	0.571	0.562
Standard LMMs									
Qwen3-VL-Plus	0.366	0.893	0.519	0.910	0.411	0.566	0.638	0.652	0.542
GPT-5.2	0.347	0.933	0.506	0.929	0.332	0.489	0.638	0.633	0.498
Claude-4.5-opus	0.309	**0.995**	0.472	**0.988**	0.148	0.257	0.648	0.571	0.364
Reasoning-Enhanced LMMs (Long-CoT)									
Qwen3-VL-8B-Think (backbone)	0.265	0.115	0.160	0.721	**0.892**	0.797	0.493	0.504	0.479
Qwen3-VL-Plus (reasoning)	0.316	0.905	0.468	0.872	0.247	0.385	0.594	0.576	0.427
GPT-5.2 (reasoning)	0.401	0.903	0.555	0.928	0.483	0.635	0.665	0.693	0.595
Gemini-3.0-Pro (reasoning)	0.313	0.978	0.474	0.954	0.176	0.297	0.634	0.577	0.385
Claude-4.5-opus (reasoning)	0.364	0.964	0.528	0.962	0.353	0.517	0.663	0.658	0.522
Flagship Models with Social-CoT Simulation									
Gemini-2.5-Flash (social-CoT)	0.353	0.629	0.452	0.779	0.615	0.687	0.566	0.622	0.570
Qwen3-VL-Plus (social-CoT)	0.380	0.766	0.508	0.853	0.521	0.647	0.617	0.644	0.578
GPT-5.2 (social-CoT)	0.442	0.304	0.360	0.762	0.853	0.805	0.602	0.578	0.582
Claude-4.5-opus (social-CoT)	0.371	0.810	0.510	0.867	0.474	0.613	0.619	0.642	0.561
MEDEA (Ours)	**0.603**	0.705	**0.650**	0.850	0.845	**0.847**	**0.727**	**0.775**	**0.749**

Table 2: Main results on CASTER-Bench. We compare MEDEA against four categories of baselines: Traditional VQA, Standard LMMs, Reasoning-Enhanced LMMs (Long-CoT), and Social-CoT simulated models. We report precision, recall, and F1-score for the High-Quality and Low-Quality classes, as well as macro-averaged metrics. Since the CASTER task focuses on identifying truly high-quality content from high-exposure UGC, performance on the High-Quality class is particularly critical. Token overhead and reasoning cost are presented in Appendix C.

summarize key components. For unlabelled UGC items, we query Gemini-2.5-Flash to generate reasoning traces and pseudo-labels. The model receives multimodal and metadata-rich inputs, including Cover image, 7 key frames sampled from the video, Title, Tags, ASR transcript, Primary category label, Secondary category label, Video duration, Resolution, Orientation (vertical / non-vertical) and Top 50 most-liked comments from which 15–20 content-relevant comments are selected. This process yields 54k Gemini-labeled CoT samples. For the 3k human-annotated UGC items, we additionally supply the ground-truth quality label when prompting Gemini, enabling it to generate supervision traces aligned with human judgment. Prompt templates used for CoT generation are provided in Appendix F. During SFT, we train MEDEA on the combined Gemini-labeled and human-annotated corpus. During RL, we only use the human-curated samples, ensuring that the reinforcement signal is anchored to expert-quality annotations. Additional training configurations and hyperparameters are also included in Appendix D.

4.2 Main Results

Table 2 presents the main results on CASTER-Bench. A defining property of this benchmark is its imbalanced label distribution: High-Quality UGC constitutes only a small fraction of the data. Consequently, performance on the High-Quality class is the most critical metric, as it reflects a model's ability to recognize intrinsic excellence rather than merely filtering out obvious failures.

MEDEA demonstrates superior performance, significantly outperforming all baselines across every category. It achieves an F1 score of 0.650 on the High-Quality class, surpassing the strongest baseline by a large margin. Crucially, MEDEA strikes an optimal balance between precision (0.603) and recall (0.705). This indicates strong selectivity—a capability essential for practical recommendation systems where false positives degrade user trust.

Analyzing the baseline categories reveals distinct failure modes:

Generosity Bias in LMMs: A striking phenomenon is observed in both Standard LMMs and Reasoning-Enhanced LMMs. Flagship models

Method	High-Quality			Low-Quality			Macro Average		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
SFT-pseudo-label	0.379	0.683	0.487	0.802	0.599	0.686	0.591	0.641	0.587
SFT-human-label	0.341	0.407	0.371	0.755	0.670	0.710	0.548	0.539	0.541
SFT-w/o-social-CoT	0.377	0.787	0.510	0.833	0.517	0.638	0.605	0.652	0.574
SFT-pseudo+human	0.398	0.750	0.520	0.829	0.557	0.666	0.614	0.654	0.593
RL-pseudo+human	0.578	0.500	0.536	0.790	0.916	0.848	0.684	0.708	0.692
RL-w/o-social-reward	0.575	0.657	0.613	0.836	0.837	0.836	0.706	0.747	0.725
RL-w/o-social-CoT	0.504	0.362	0.421	0.770	0.880	0.821	0.637	0.621	0.621
MEDEA(RL-human-label)	0.603	0.705	0.650	0.850	0.845	0.847	0.727	0.775	0.749

Table 3: Ablation studies on CASTER-Bench. Each component of MEDEA contributes to overall performance.

like GPT-5.2 and Claude-4.5-Opus achieve near-perfect Recall ($> 90\%$) on the High-Quality class but suffer from extremely low Precision ($\sim 30\%$). This suggests that while these models can identify positive attributes in almost any video via long-context reasoning, they exhibit a "Generosity Bias". They tend to over-rationalize merit in average content, lacking the critical social discernment to distinguish "acceptable" content from "community-resonant" masterpieces.

Signal-Dominance in Traditional VQA: Traditional methods (e.g., FastVQA, VQA2) are heavily biased towards Low-Quality class. Their High-Quality F1 scores remain consistently poor (ranging from 0.33 to 0.41), confirming that aesthetic fidelity alone is insufficient for capturing the semantic and social dimensions of community resonance.

Effectiveness of Social Alignment: While prompting flagship models with Social-CoT (the fourth category) improves performance over standard zero-shot inference, they still lag behind MEDEA. For instance, Qwen3-VL-Plus with Social-CoT achieves an F1 of 0.508 compared to MEDEA's 0.650. This validates that the reasoning pattern alone is not enough; the model requires the specific alignment with expert-curated social judgments provided by MEDEA's training pipeline to internalize the true "community standard".

Finally, MEDEA achieves the highest Macro-F1 score (0.749), reflecting robust performance across the entire quality spectrum. Its ability to maintain high recall without succumbing to the positivity bias of general-purpose reasoning models validates the effectiveness of the proposed framework.

4.3 Ablation Experiments

To isolate the contribution of each component in MEDEA, we perform a series of ablations. Specifically, we analyze the impact of the Social-CoT and the Social Alignment Reward (denoted as r_{social}).

Necessity of Social Reasoning Paths. Removing the Social-CoT ("RL-w/o-social-CoT") leads to a substantial performance drop, with the High-Quality F1 score decreasing from 0.650 to 0.421. This sharp decline confirms that pixel-level perception alone is insufficient for assessing community resonance. The Social-CoT acts as a necessary cognitive bridge, allowing the model to perform multimodal perspective-taking to infer how content features translate into user engagement.

Impact of Social Alignment and Qualitative Analysis. Excluding the Social Alignment Reward leads to *Social Mode Collapse*, where reasoning degenerates into repetitive, generic templates (e.g., "So beautiful"). Qualitative inspection (Appendix E) confirms this distinction: while MEDEA empathetically interprets wind-swept keyframes in an Iceland vlog as "raw natural power", the ablated model produces only hollow praise. This underscores that social alignment is critical for grounding the model in authentic, emotionally nuanced community expression.

5 Conclusions

This work establishes a new paradigm for UGC assessment, shifting focus from aesthetic fidelity to social-cognitive resonance. By introducing the Social-CoT mechanism, we demonstrate that effective quality assessment requires not just signal analysis, but the capacity for multimodal perspective-taking. Our framework, MEDEA, validates that simulating a "community of critics" via Social Alignment Reward effectively captures the nuance of human engagement. Beyond specific performance gains on CASTER-Bench, this research paves the way for equipping LMMs with Theory of Mind capabilities, bridging the gap between computational metrics and genuine social understanding.

Limitations

While MEDEA demonstrates strong performance on community-aware UGC assessment, several limitations remain. First, although the Social-CoT mechanism incurs additional computational cost compared to direct prediction (as detailed in Appendix C), this overhead is slightly higher than that of some reasoning-enhanced LMMs, but since MEDEA has a much smaller parameter size, the overall cost and inference time remain controllable. Second, the current social alignment is optimized for specific platform dynamics; consequently, its generalizability to other social ecosystems with distinct cultural norms or community behaviors remains to be verified. Third, our binary framing oversimplifies the continuous spectrum of community resonance. Finally, while our current implementation leverages rich multimodal metadata for social grounding, the MEDEA framework is theoretically extensible to single-modality or sparse-signal scenarios, which we leave for future exploration.

References

Duolikun Danier, Fan Zhang, and David R Bull. 2023. Bvi-vfi: A video quality database for video frame interpolation. *IEEE Transactions on Image Processing*, 32:6004–6019.

Axel De Decker, Jan De Cock, Peter Lambert, and Glenn Van Wallendael. 2024. No-reference vmaf: A deep neural network-based approach to blind video quality assessment. *IEEE Transactions on Broadcasting*, 70(3):844–861.

Huiyu Duan, Qiang Hu, Jiarui Wang, Liu Yang, Zitong Xu, Lu Liu, Xiongkuo Min, Chunlei Cai, Tianxiao Ye, Xiaoyun Zhang, and 1 others. 2025. Finevq: Fine-grained user generated content video quality assessment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3206–3217.

Qihang Ge, Wei Sun, Yu Zhang, Yunhao Li, Zhongpeng Ji, Fengyu Sun, Shangling Jui, Xiongkuo Min, and Guangtao Zhai. 2025. Lmm-vqa: Advancing video quality assessment with large multimodal models. *IEEE Transactions on Circuits and Systems for Video Technology*.

Chenlong He, Qi Zheng, Ruoxi Zhu, Xiaoyang Zeng, Yibo Fan, and Zhengzhong Tu. 2024. Cover: A comprehensive video quality evaluator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5799–5809.

Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe. 2017. The konstanz natural video database (konvid-1k). In *2017 Ninth international conference on quality of multimedia experience (QoMEX)*, pages 1–6. IEEE.

Ziheng Jia, Zicheng Zhang, Jiaying Qian, Haoning Wu, Wei Sun, Chunyi Li, Xiaohong Liu, Weisi Lin, Guangtao Zhai, and Xiongkuo Min. 2025. Vqa2: visual question answering for video quality assessment. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 6751–6760.

Jari Korhonen. 2019. Two-level approach for no-reference consumer video quality assessment. *IEEE Transactions on Image Processing*, 28(12):5923–5938.

Dingquan Li, Tingting Jiang, and Ming Jiang. 2019. Quality assessment of in-the-wild videos. In *Proceedings of the 27th ACM international conference on multimedia*, pages 2351–2359.

Guo Li, Baoliang Chen, Lingyu Zhu, Qinwen He, Hongfei Fan, and Shiqi Wang. 2021. Pugcq: A large scale dataset for quality assessment of professional user-generated content. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3728–3736.

Joe Yuchieh Lin, Rui Song, Chi-Hao Wu, TsungJung Liu, Haiqiang Wang, and C-C Jay Kuo. 2015. Mcl-v: A streaming video quality assessment database. *Journal of Visual Communication and Image Representation*, 30:1–9.

Yiting Lu, Xin Li, Yajing Pei, Kun Yuan, Qizhi Xie, Yunpeng Qu, Ming Sun, Chao Zhou, and Zhibo Chen. 2024. Kvq: Kwai video quality assessment for short-form videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25963–25973.

Alex Mackin, Fan Zhang, and David R Bull. 2019. A study of high frame rate video formats. *IEEE Transactions on Multimedia*, 21(6):1499–1512.

Pavan C Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C Bovik. 2021a. St-greed: Space-time generalized entropic differences for frame rate dependent video quality prediction. *IEEE Transactions on Image Processing*, 30:7446–7457.

Pavan C Madhusudana, Xiangxu Yu, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C Bovik. 2021b. Subjective and objective quality assessment of high frame rate videos. *IEEE Access*, 9:108069–108082.

K Manasa and Sumohana S Channappayya. 2016. An optical flow-based full reference video quality assessment algorithm. *IEEE Transactions on Image Processing*, 25(6):2480–2492.

Anush Krishna Moorthy and Alan Conrad Bovik. 2010. Efficient video quality assessment along temporal trajectories. *IEEE transactions on circuits and systems for video technology*, 20(11):1653–1658.

Mikko Nuutinen, Toni Virtanen, Mikko Vaahteranoksa, Tero Vuori, Pirkko Oittinen, and Jukka Häkkinen. 2016. Cvd2014—a database for evaluating no-reference video quality assessment algorithms. *IEEE Transactions on Image Processing*, 25(7):3073–3086.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of large language models when interaction requires anticipating others' states. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8184–8205.

Kalpana Seshadrinathan, Rajiv Soundararajan, Alan Conrad Bovik, and Lawrence K Cormack. 2010. Study of subjective and objective quality assessment of video. *IEEE transactions on Image Processing*, 19(6):1427–1441.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Zeina Sinno and Alan Conrad Bovik. 2019. Large-scale study of perceptual video quality. *IEEE Transactions on Image Processing*, 28(2):612–627.

Tingyu Song, Tongyan Hu, Guo Gan, and Yilun Zhao. 2025. Vf-eval: Evaluating multimodal llms for generating feedback on aigc videos. *arXiv preprint arXiv:2505.23693*.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, and 1 others. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.

Domonkos Varga. 2022. No-reference video quality assessment using multi-pooled, saliency weighted deep features and decision fusion. *Sensors*, 22(6):2209.

Phong V Vu, Cuong T Vu, and Damon M Chandler. 2011. A spatiotemporal most-apparent-distortion model for video quality assessment. In *2011 18th IEEE international conference on image processing*, pages 2505–2508. IEEE.

Xinyi Wang, Angeliki Katsenou, Junxiao Shen, and David Bull. 2025. Camp-vqa: Caption-embedded multimodal perception for no-reference quality assessment of compressed video. *arXiv preprint arXiv:2511.07290*.

Yilin Wang, Sasi Inguva, and Balu Adsumilli. 2019. Youtube ugc dataset for video compression research. In *2019 IEEE 21st international workshop on multimedia signal processing (MMSP)*, pages 1–5. IEEE.

Zhou Wang and Qiang Li. 2007. Video quality assessment using a statistical model of human visual speed perception. *Journal of the optical society of america A*, 24(12):B61–B69.

Zhou Wang, Hamid R Sheikh, and Alan C Bovik. 2002. No-reference perceptual quality assessment of jpeg compressed images. In *Proceedings. International conference on image processing*, volume 1, pages I–I. IEEE.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. 2022. Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling. *Proceedings of European Conference of Computer Vision (ECCV)*.

Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. 2023a. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20144–20154.

Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. 2023b. Towards explainable in-the-wild video quality assessment: a database and a language-prompted approach. In *Proceedings of the 31st acm international conference on multimedia*, pages 1045–1054.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, and 1 others. 2024. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. In *International Conference on Machine Learning*, pages 54015–54029. PMLR.

Fengchuang Xing, Yuan-Gen Wang, Hanpin Wang, Leida Li, and Guopu Zhu. 2022. Starvqa: Space-time attention for video quality assessment. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2326–2330. IEEE.

Zicheng Zhang, Wei Wu, Wei Sun, Danyang Tu, Wei Lu, Xiongkuo Min, Ying Chen, and Guangtao Zhai. 2023. Md-vqa: Multi-dimensional quality assessment for

ugc live videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1746–1755.

A Related Works

A.1 UGC Databases

Early UGC benchmarks (Seshadrinathan et al., 2010; Lin et al., 2015; Nuutinen et al., 2016; Mackin et al., 2019; Madhusudana et al., 2021b; Danier et al., 2023) mainly relied on professionally produced videos with controlled, synthetic distortions. Recent datasets have shifted focus toward authentic, in-the-wild UGC with large-scale crowdsourced annotations, including KoNViD-1k (Hosu et al., 2017), LIVE-VQC (Sinno and Bovik, 2019), YouTube-UGC (Wang et al., 2019), and PUGCQ (Li et al., 2021), which better reflect real-world content diversity and mixed distortions.

Beyond overall quality scores, recent efforts have moved toward multi-dimensional quality modeling by disentangling aesthetic and technical factors. Notable examples include datasets explored in DOVER (Wu et al., 2023a), MD-VQA (Zhang et al., 2023), MaxVQA (Wu et al., 2023b), KVQ (Lu et al., 2024), and FineVQ (Duan et al., 2025). In parallel, VF-EVAL (Song et al., 2025) introduces a benchmark for evaluating MLLMs' ability to generate feedback on AIGC videos, focusing on prompt alignment, coherence, and commonsense reasoning. However, these datasets and benchmarks predominantly emphasize perceptual attributes or feedback correctness for short-form or synthetic videos. In contrast, CASTER-Bench targets long-form, real-world UGC and explicitly models social-cognitive judgments such as narrative engagement and emotional resonance, which are critical for understanding community-level content appreciation.

A.2 UGC-VQA Models

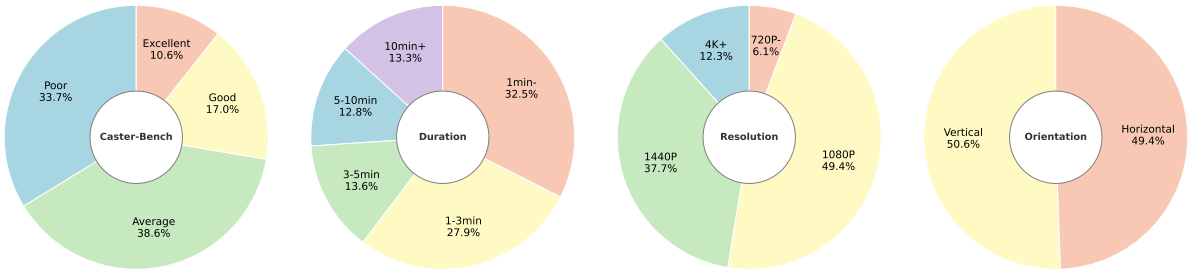
UGC-VQA methods have evolved from full-reference metrics (Manasa and Channappayya, 2016; Wang and Li, 2007; Moorthy and Bovik, 2010; Madhusudana et al., 2021a; Vu et al., 2011), which require unavailable references, to no-reference approaches. Classical models leveraged handcrafted statistical priors (Wang et al., 2002; Korhonen, 2019), while modern approaches learn content-dependent spatiotemporal representations from large-scale distorted data (Varga, 2022; De Decker et al., 2024; Li et al., 2019; Zhang et al., 2023; Xing et al., 2022; Ge et al., 2025; Duan et al., 2025). Representative methods include VSFA (Li et al., 2019) (temporal modeling), MD-VQA (Zhang et al., 2023) (fusion of spatial, motion, and semantic cues), StarVQA (Xing et al., 2022) (self-attention on salient spatiotemporal regions), and DOVER (Wu et al., 2023a) (dual-branch modeling of technical quality and aesthetic preference).

The recent advent of vision-language pretraining has catalyzed multimodal directions in UGC-VQA (Radford et al., 2021; Tschannen et al., 2025). CLIP-based methods, such as COVER (He et al., 2024) and MaxVQA (Wu et al., 2023b), employ semantic encoders to inject high-level content priors. Furthermore, prompt-driven alignment methods like Q-Align (Wu et al., 2024) enable zero-shot or cross-modal approximation of human judgments. Emerging Large Multimodal Models (LMMs), such as LMM-VQA (Ge et al., 2025), FineVQ (Duan et al., 2025), and CAMP-VQA (Wang et al., 2025), integrate spatial, temporal, and text-based reasoning to produce robust quality estimates. However, these methods typically treat text as a static feature rather than utilizing it to simulate the dynamic social reception of the content.

A.3 Chain-of-Thought and Social Intelligence

While Chain-of-Thought (CoT) prompting has revolutionized large language model performance in logical, mathematical, and symbolic reasoning tasks (Wei et al., 2022; Shao et al., 2024), its application to social intelligence remains a frontier challenge. Recent studies in Theory of Mind (ToM) investigate whether LLMs can effectively infer the mental states, beliefs, and emotional reactions of others (Sap et al., 2022). In the context of UGC assessment, we argue that quality is not an intrinsic property of the signal but a product of social reception.

Our work bridges these domains by proposing Social-CoT. Unlike standard CoT which focuses on step-by-step logical deduction, Social-CoT explicitly operationalizes ToM by simulating diverse viewer personas and their empathetic engagement paths. This approach shifts the evaluation paradigm from analyzing static content features to simulating the "community mind", thereby aligning computational quality assessment with authentic community dynamics.



(a) UGC Resonance Grades (b) Duration (c) Resolution (d) Orientation

Figure 3: Distribution across various dimensions in CASTER-Bench. (a) Proportion of UGC Resonance annotated by experts. (b)-(d) Proportion of UGC items static features including duration, resolution and orientation.

B Data Collection and Statistics

The technical attributes of CASTER-Bench follow natural distributions consistent with real-world UGC environments, as shown in Figure 3.

- **Duration**: Unlike traditional datasets focused on short clips (8-10s), our benchmark contains a significant portion of long-form videos, with an average duration of 442 seconds. We observe that video duration correlates more strongly with perceived quality than technical attributes.
- **Resolution**: The dataset covers a wide range of resolutions, reflecting the diverse recording devices used by creators.
- **Orientation**: The split between vertical and horizontal videos is naturally balanced, catering to both mobile and desktop viewing habits.

C Token Overhead and Reasoning Cost

Table 4 details the computational overhead associated with the reasoning process. Integrating the Social-CoT module significantly increases the generation volume: MEDEA generates an average of 1,256 tokens per UGC item, compared to just 5.6 tokens for the direct-answer variant (MEDEA w/o Social-CoT).

Inference Efficiency. We evaluate efficiency on local 4×H800 GPUs using vLLM with 8 concurrent workers. The generation of dense social reasoning reduces inference throughput from 2.55 to 0.79 videos/sec. However, this increased latency is a necessary trade-off for precision. As shown in Table 2, this computational investment yields a High-Quality F1 score of 0.650, outperforming the fastest traditional methods (F1 $\approx$ 0.33–0.41) which fail to capture semantic resonance.

Comparison with Reasoning Baselines. Analyzing the relationship between token consumption and performance reveals that simply increasing reasoning length does not guarantee better judgment:

- **Inefficient Deep Reasoning:** High token consumption does not automatically translate to high accuracy. For instance, Qwen3-VL-Plus (reasoning) generates nearly 1,000 tokens per video (917.5) but only achieves a High-Quality F1 of 0.468. Despite a reasoning depth comparable to ours, it lacks the specific social alignment, resulting in verbose but ultimately misaligned judgments that succumb to the generosity bias.
- **Shallow Reasoning Limits:** Conversely, models with lower reasoning overheads, such as GPT-5.2 (reasoning) and Gemini-3.0-Pro (reasoning), consume significantly fewer tokens (96.5 and 160.0, respectively). However, this efficiency caps their performance (High-Quality F1 of 0.555 and 0.474), suggesting that the complex social dynamics of UGC cannot be adequately captured through brief, surface-level chain-of-thought processes.
- **Simulation vs. Alignment:** Flagship models prompted with Social-CoT (e.g., Claude-4.5-Opus at 712.4 tokens) sit in the middle ground, utilizing moderate token budgets to simulate social critique. Yet, they still fall short of MEDEA (F1 0.510 vs. 0.650). This indicates that MEDEA's higher token count (1,256) is not merely verbose, but represents a necessary depth of analysis derived from training on expert data—internalizing a standard that prompt engineering alone cannot fully replicate.

Method	Videos/sec	Tokens
Ours (Local Inference)		
MEDEA-w/o-social-CoT	2.55	5.6
MEDEA	0.79	1,256.0
Reasoning-Enhanced LMMs (API-based)		
GPT-5.2 (reasoning)	-	96.5
Gemini-3.0-Pro (reasoning)	-	160.0
Claude-4.5-Opus (reasoning)	-	563.4
Qwen3-VL-Plus (reasoning)	-	917.5
Social-CoT Simulation (API-based)		
Qwen3-VL-Plus (social-CoT)	-	424.9
GPT-5.2 (social-CoT)	-	489.2
Claude-4.5-Opus (social-CoT)	-	712.4

Table 4: Average tokens per UGC item and inference efficiency. MEDEA's higher token count reflects the generation of dense social context, which is critical for High-Quality identification. Baselines are API-based; speed/hardware not reported.

Stage	Hyperparameter	Value
SFT	batch size	256
	learning rate	5e-6
	learning rate schedule	cosine
	learning rate decay ratio	0.2
RL	batch size	64
	learning rate	1e-6
	learning rate schedule	cosine
	learning rate decay ratio	0.1
	PPO clip ratio low	0.2
	PPO clip ratio high	0.2
	kl coefficient	0.001
	entropy coefficient	0.001
	rollout number	8
	rollout top-p	1.0
	rollout temperature	0.6
	rollout repetition penalty	1.0
Inference	top-k	50
	top-p	0.7
	temperature	0.6
	repetition penalty	1.0

Table 5: Hyperparameters used in training and inference of MEDEA.

In summary, MEDEA leverages a higher token budget to construct a critical social context that other models either gloss over (shallow reasoners) or misinterpret through excessive positivity (deep reasoners).

D Hyperparameters used in training and inference of MEDEA

Hyperparameters used in training and inference of MEDEA are presented in Table 5.

E Qualitative Analysis of Social Reasoning Paths

To qualitatively illustrate how the Social-CoT mechanism instantiates diverse viewer personas to achieve social reasoning, we present representative examples of reasoning paths under three settings:

1. **Oracle Social Context** (Figure 6): Social-CoT generated by a strong proprietary model (Gemini) conditioned on *real, high-engagement user comments*. This serves as the "upper bound" or gold standard for community-aligned reasoning.

2. **Social-CoT with Alignment** (Figure 7): Reasoning paths generated by MEDEA using the full Social Alignment Reward (r_{social}). This demonstrates the model's capability for *Empathetic Simulation*.

3. **Social-CoT without Alignment** (Figure 8): Reasoning paths generated by MEDEA without the social alignment constraint. This illustrates the phenomenon of "Social Mode Collapse", where reasoning becomes repetitive and robotic.

We additionally provide the UGC item cover image together with seven uniformly sampled key frames in Figure 4, which serve as the visual context available to the model during perspective-taking. These frames capture representative scenes, visual quality, and narrative progression, enabling readers to assess how well the generated Social-CoT aligns with the visual narrative.

Analysis of Oracle Social Context. The first setting (Figure 6) serves as a reference for authentic social cognition. By accessing real community feedback, the reasoning path exhibits rich, fine-grained emotional granularity, connecting specific visual metaphors (e.g., "The wilderness is a determination") to deep philosophical reflections found in the comment section.

Analysis of Social Alignment. The comparison between MEDEA with and without Social Alignment highlights the emergence of social intelligence.

As illustrated in the case study of an Iceland trip vlog (Figure 7), MEDEA demonstrates the ability to simulate empathy. Instead of merely listing technical attributes like resolution or lighting, the model instantiates diverse viewer personas to evaluate the content's visual narrative. For instance, by analyzing key frames that depict people walk-

(a) Cover (b) Key Frame 1 (c) Key Frame 2 (d) Key Frame 3

(e) Key Frame 4 (f) Key Frame 5 (g) Key Frame 6 (h) Key Frame 7

Figure 4: Cover and 7 uniformly sampled key frames of the example.

ing against strong gusts, the model interprets this not just as motion, but as a manifestation of Iceland's raw natural power. It consequently simulates a viewer's visceral reaction: "The wind in Iceland looks intense, really shocking". This indicates that MEDEA has internalized the nuanced, multi-faceted "voice" of the community.

In stark contrast, Figure 8 (Without Alignment) demonstrates Social Mode Collapse. While the model correctly identifies the content as "beautiful", the reasoning path degenerates into repetitive templates (e.g. repeating "So beautiful... I really want to go" multiple times). This confirms that without the Social Alignment Reward, the model fails to capture the diverse "voice" of the community, resulting in a hollow simulation lacking empathetic depth.

Overall, these examples demonstrate that Social-CoT can effectively substitute real user feedback in driving engagement-aware reasoning, and that the Social Alignment Reward plays a crucial role in improving the authenticity, coherence, and interpretability of the generated reasoning process.

F Prompts used in MEDEA

We present the complete prompt used to instruct the teacher model to perform comment selection, stance classification, and reasoning-based aggregation for UGC items. The prompt is designed to simulate how users infer the creative quality of a UGC item from its visual and textual content, and how such inferences are reflected in the comment section.

The task formulation explicitly constrains the model to rely only on observable video attributes, including the cover image, key frames, metadata, and automatically transcribed text, while excluding any auditory or external signals. To ensure interpretability and reproducibility, the prompt enforces strict rules on comment selection, independent coverage of each comment, and a final statistically grounded stance decision based on a Skellam-normalized difference score. The prompt used to generate reasoning content is presented in Figure 9.

We design a structured prompt to guide MEDEA in simulating comment-section reactions on UGC items. The prompt integrates both visual inputs (cover image and key frames) and textual metadata (title, tags, ASR, category, and video attributes), encouraging the model to reason about the perceived creation quality of a UGC item. Instead of directly predicting an overall label, the model is required to first generate a diverse set of stance-aware comments. The final judgment is derived through a quantitative aggregation process based on a Skellam z-score, which measures the normalized difference between supportive and opposing comments. This design enforces internal consistency, reduces shortcut learning, and aligns the prediction with interpretable intermediate reasoning. The prompt used to train MEDEA is presented in Figure 10.

G Statistical Significance Testing

To more comprehensively evaluate the performance of our MEDEA method, we incorporated p-values alongside conventional metrics in Table 6. The consistent statistical significance observed across all experimental results, as clearly demonstrated in the accompanying table, strongly attests to the robustness of our approach. These findings not only provide compelling evidence that our method substantially outperforms the baseline but also highlight its reliability and generalizability under varied conditions.

Method	Macro Average		
	Precision	Recall	F1
MEDEA	5.0e-04	3.4e-02	<1.0e-10

Table 6: P-values comparing MEDEA with the best baseline (GPT-5.2 reasoning) using paired bootstrap tests.

H Distinguishing Intrinsic Quality from Popularity

In this study, the core objective of the CASTER task is to assess the intrinsic value of UGC items, rather than merely predicting their current popularity, which is influenced by various external factors. The expert-annotated dataset we employ essentially serves as a "refinement" and "correction" of the noisy real-world community signals. Authentic user interaction data is saturated with noise, such as click-farming bots, irrational herd behavior, and biases inherent in the platform's recommendation algorithms. Therefore, expert annotations provide a well-considered and idealized signal based on the intrinsic value of the content itself.

To illustrate this point more tangibly, we present some representative cases observed in the dataset in Table 5, which demonstrate the fundamental distinction between learning from expert judgments and blindly fitting popularity metrics. Certain UGC items with high actual view counts or interaction metrics are labeled as low-quality by experts. Such content often relies on sensationalist titles, vulgar visual elements, or misleading information, with high traffic stemming more from emotional provocation or short-term platform recommendation strategies than intrinsic value.

By training models to fit this "refined" expert signal, the CASTER task aims to advance the modeling and recognition of content quality itself.

I Detailed Results of Baselines

Most of the compared baselines are originally designed as regression-based methods, which output continuous quality scores rather than discrete class labels. To ensure a fair and informative comparison under the classification setting adopted in this work, we perform threshold sweeping on the CASTER-Bench for all regression-based methods.

Specifically, for each method, we vary the decision threshold that maps predicted quality scores to discrete quality categories and evaluate the corresponding classification performance. The threshold that yields the best macro-averaged F1 score is selected and reported as the main result in the paper. This procedure allows each method to operate under its optimal decision boundary, avoiding performance degradation caused by suboptimal or arbitrary threshold choices.

We present the complete performance results of each method under different threshold settings. Detailed results for FastVQA, DOVER, MaxVQA, Q-Align, FineVQ, and VQA2 can be found in Table 7, Table 8, Table 9, Table 10, Table 11, and Table 12, respectively.

J Declaration of AI Assistance

We utilized Gemini to refine the wording and correct grammatical errors in the drafting of this paper. The authors reviewed and revised all AI-generated suggestions to ensure accuracy and consistency with the original ideas.

Cases of High-Popularity but Low-Expert-Rated Content

Example1

Title:
The Unboxing of WLOP's Art Collection "Ghostblade 4" Is Here!! **There's a Giveaway**!
Tag:
Unboxing, Sharing, Ghostblade, Artbook, WLOP, Ghostblade 4
ASR:
WLOP's new art collection Ghostblade 4 is out!! Here we're giving away a brand new artbook to our lovely friends! **You'll need at least 80 likes and 10 comments to participate**. The lottery draw will be held at 7 PM on August 29! Everyone, don't forget to like, comment, and share! Features newly drawn illustrations from the Ghostblade and Cloud Insect series created between 2022 and 2025 . . . . . .
Comments:
"So stunning! [Sparkling eyes] And please pick me for the giveaway! [Grin] [Smirking]"
"I've always adored WLOP's art style [Cloud Zoo Hot Air Balloon]. Hearing you say the Ghostblade 4 collector's edition has amazing quality and is super-sized has totally raised my expectations. I really want to feel that texture in my own hands!"
"Wow, this is so exquisite!" . . . . . .
"Love it! [Cheering]"

Example2

Title:
Have You Slept Well, Puppy?
Tag:
Sleep, Sleep aid, Hypnosis, Livestream, Streamer
ASR:
(BGM) Wanna be in it forever girl, you know I adore ya . . . . . .
Comments:
"I love this angle, it's perfect! [heart]"
"Wow wow wow [cheering][cheering][cheering]. I can't fall asleep now! "
"Love it [heart]"
"I really like this one."
"I like you [giving flowers]"

Figure 5: Representative examples of "inflated bubbles": videos with high popularity metrics that experts rated as low-quality. Although user comments all show positive sentiment, experts have determined that these videos contain giveaway incentives and guided commenting behaviors, or include suggestive content, and therefore believe the video quality should be rated as low-quality.

threshold	High-Quality			Low-Quality			Macro Average		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
0.055	0.277	**1.000**	0.434	0.000	0.000	0.000	0.138	0.500	0.217
0.065	0.277	**1.000**	0.434	**1.000**	0.001	0.002	**0.639**	0.500	0.218
0.275	0.277	0.966	0.431	0.725	0.034	0.066	0.501	0.500	0.248
0.485	0.292	0.903	**0.441**	0.812	0.161	0.269	0.552	0.532	0.355
0.695	0.326	0.676	0.440	0.790	0.465	0.585	0.558	**0.571**	0.513
0.795*	0.347	0.440	0.388	0.761	0.682	0.719	0.554	0.561	**0.554**
0.915	**0.500**	0.002	0.005	0.723	0.999	**0.839**	0.612	0.501	0.422
0.925	0.000	0.000	0.000	0.723	**1.000**	**0.839**	0.362	0.500	0.420

Table 7: Performance comparison using FastVQA under different thresholds. Best threshold is marked with $\star$, and best results in each column are highlighted in bold.

Oracle Social Context: CoT generated by Gemini conditioned on real user comments

Based on the provided video metadata, this video belongs to the "Travel & Outdoors" primary category and the "Travel Vlog" subcategory. The title, "Studying abroad or going overseas cannot change my life," directly highlights the core reflection of the video. The video duration is "288 seconds," with a resolution of "3840×2160," and it is not in vertical format, indicating relatively high visual quality. The ASR content reveals deeper philosophical reflections, covering travel experiences in "Iceland," a redefinition of the concept of the "wilderness," discussions on personal "choice and responsibility," and references to and interpretations of the "Frame 25" metaphor from the film *The Secret Life of Walter Mitty*, ultimately emphasizing the importance of "determination." The cover image text, "The wilderness is merely a determination," strongly aligns with the title and ASR content, jointly establishing a contemplative tone. Key frames showcase Icelandic natural landscapes, urban scenery, road trips, and moments of personal reflection by the creator, resulting in visually appealing and narratively rich visuals.

When watching videos that integrate travel documentation, personal reflection, and philosophical exploration, audiences typically focus on several aspects: the depth of the content and the emotional resonance it evokes, the creator's unique perspectives on life and decision-making, the aesthetic quality and narrative coherence of the travel footage, and whether the video provides emotional comfort or new ways of thinking.

When viewers encounter the ASR content describing the creator's personal experiences and reflections on "studying abroad," as well as discussions of "courage" and "insight," audiences with similar life experiences tend to develop a strong sense of identification and resonance. **["I am a 25-fall freshman at a university in Singapore. This video gave me a lot of reflection, insight, and courage. Loved it [applause]"]** (supportive comment 1)

When viewers resonate with the ASR's in-depth discussion of "travel" and the delicate portrayal of the "withdrawal" state after returning from a trip, especially those who have just come back from traveling, they are deeply moved by how accurately the video captures their inner feelings. **[[cry][cry][cry] I cried watching this. I just returned from Southern Europe yesterday and have been stuck in a withdrawal state. I happened to see this video on my feed. Thank you so much—it spoke directly to my heart.]** (supportive comment 2)

The ASR content deeply explores the core arguments of "the wilderness as determination" and "taking responsibility for one's choices," prompting some viewers to engage in profound philosophical reflection and extend the video's viewpoints through their own experiences. **[Long philosophical reflection elaborating on anxiety, responsibility, action, freedom, and the meaning of "the wilderness," concluding that making a firm decision to confront life's difficulties places one already within the wilderness.]** (supportive comment 3)

When viewers encounter the ASR's explicit statement that "we are not free when we cannot take responsibility for what lies ahead or for our own choices," along with its interpretation of "anxiety," they feel understood and emotionally healed. **[Thank you, uploader! I keep rewatching this, and I really feel healed while being anxious [cry][cry]!! When I cannot take responsibility for what is in front of me or for my own choices, I am not free. . .]** (supportive comment 4)

.....

The ASR's firm declaration that "a decision is the best decision" and "taking responsibility for one's choices" strongly motivates viewers who are at a low point in life. **[This is so well written. It really encouraged me when I was at rock bottom.]** (supportive comment 13)

When viewers encounter this philosophically rich "travel vlog" during their own travels, the alignment between the content and their current situation creates especially strong resonance. **[I saw this video while living in Dali during my travels, and it really moved me.]** (supportive comment 14)

Regarding the discussion of "freedom" and "value" implied by the title "Studying abroad or going overseas cannot change my life," some viewers adopt a critical stance, believing the creator failed to extract higher value from their experiences. **[Your freedom did not bring you greater value.]** (opposing comment 1)

When the cover and ASR present "the wilderness as determination" as the core idea, some viewers perceive this insight as unoriginal or superficial and respond with sarcasm. **[That just means you haven't matured mentally. If the wilderness is determination, then I already had it when I was six or seven.]** (opposing comment 2)

Given the high resolution, refined key frames, and depth of thought expressed in the ASR, viewers feel the video's high production quality deserves more attention and lament its relatively low view count. **[This video is amazing! How does it only have 37k followers?!]** (supportive comment 15)

The ASR's final reveal of "Frame 25" and the creator's promise for the future provide profound insight and emotional elevation, leading viewers to regard the ending as exceptionally well-crafted. **[The ending is brilliant.]** (supportive comment 16)

The ASR's deep interpretation of "what the wilderness is" and "paying the price for one's actions" offers strong emotional support and meaningful guidance to viewers facing life difficulties. **[I really love this so much. You have no idea how important this was to me while I was going through a hard time. Especially the explanation of what the wilderness means and taking responsibility for one's actions—it strengthened my sense of meaning.]** (supportive comment 17)

In the simulated comments, there are **17 supportive comments** and **2 opposing comments**. The Sigma-normalized difference (Skellam z-score) is computed as

$$z = \frac{17 - 2}{\sqrt{17 + 2}} = \frac{15}{\sqrt{19}} \approx 3.441.$$

Since $z \geq 1.5$, the overall stance is classified as **supportive**.

Figure 6: Oracle Social Context: Social-CoT reasoning path generated by Gemini, grounded in real high-engagement user comments. This represents the gold standard for social reasoning.

Social-CoT generated by MEDEA with Social Alignment Reward

Based on the provided script information, the video belongs to the *Travel & Outdoors* category, specifically the *Travel Log* genre. The title, *"Iceland Vlog! One Day You Will Come to the End of the World,"* clearly establishes both the theme and emotional tone, framing the video as a documentary-style record of an Icelandic journey and emphasizing Iceland's image as the "end of the world."
The video has a duration of **351 seconds**, a high resolution of **3840×2160**, and is presented in a non-vertical format, indicating a high level of production quality with a strong emphasis on visual experience. The ASR transcript provides a detailed narration of various locations in Iceland, including Reykjavík, the Snæfellsnes Peninsula, black sand beaches, and Vatnajökull Glacier. It describes Iceland's natural landscapes, cultural elements (e.g., colorful houses and Christmas atmosphere), and travel experiences (e.g., extreme winds and safety considerations), while also conveying reflective insights such as "capturing a rainbow in the eye of the storm."
The keyframes further showcase iconic Icelandic scenes, including a red-roofed church, a black church, black sand beaches, ice blocks, and sunsets, reinforcing the video's comprehensive visual portrayal of Iceland's scenery.
Viewers watching high-quality travel vlogs of this kind typically focus on several aspects. First, they attend to visual presentation and editing quality, assessing whether the destination's characteristics are effectively showcased. Second, they evaluate content richness and depth, including whether the video offers practical travel information or meaningful reflections. Third, they consider the creator's narrative style and expressiveness, and whether it resonates emotionally. Finally, such videos often stimulate viewers' curiosity, admiration, and desire to explore a unique destination like Iceland.
When viewers observe that *the title explicitly frames the content as an Iceland travel vlog* and that *the ASR provides detailed descriptions of diverse locations and experiences*, they tend to appreciate the richness of the content. **["Iceland is so beautiful, and the vlog is very rich in content."]** (Supportive Comment 1)
Given the *4K resolution (3840×2160) and the keyframes displaying Iceland's magnificent landscapes*, viewers are impressed by the visual clarity and image quality. **["Iceland is stunning, and the visuals are incredibly sharp."]** (Supportive Comment 2)
Based on the *vivid ASR descriptions of Iceland's scenery, such as the coexistence of an "end-of-the-world" atmosphere and "soft pink gentleness"*, viewers develop a deeper understanding of Iceland's unique charm. **["I didn't know Iceland was like this—it feels both apocalyptic and gently beautiful."]** (Supportive Comment 3)
Considering the *compelling presentation of Iceland throughout the video*, viewers express a desire to see more content about this destination. **["So beautiful—please film more about Iceland."]** (Supportive Comment 4)
From the *creator's insights and reflections conveyed in the ASR*, viewers perceive depth beyond surface-level scenery. **["This is impressive—you really understand Iceland instead of just filming the scenery."]** (Supportive Comment 5)
In response to the *ASR's description of extreme winds and keyframes showing people walking against strong gusts*, viewers are struck by Iceland's raw natural power. **["The wind in Iceland looks intense—really shocking."]** (Supportive Comment 6)
Based on the *shared travel experiences in the ASR*, viewers admire the creator's extensive travel background. **["Amazing—it feels like you've traveled all over the world."]** (Supportive Comment 7)
Some viewers, drawing on their *knowledge of Iceland*, offer mild suggestions while remaining supportive overall. **["This is great, though Iceland also has hot springs and auroras that weren't fully covered."]** (Supportive Comment 8)
Viewers respond positively to the *ASR's depiction of southern Iceland*, expressing particular appreciation for that region. **["This is impressive—I especially love the scenery in southern Iceland."]** (Supportive Comment 9)
Those with *personal travel experience in Iceland* find the ASR's descriptions highly accurate. **["I went to Iceland last year, and the video describes it perfectly."]** (Supportive Comment 10)
From the *overall production quality*, viewers recognize the creator's dedication and effort. **["This video is extremely well-made—you can really feel the effort."]** (Supportive Comment 11)
Some viewers supplement the *ASR's travel advice* with personal experiences, while maintaining a supportive stance. **["I've been to Iceland too—it's not the easiest place to travel, but the video gives good warnings and is very well done."]** (Supportive Comment 12)
Based on the *ASR's emphasis on Iceland's colors*, viewers praise the video's visual expressiveness. **["Iceland looks so colorful in this video—it's beautifully shot."]** (Supportive Comment 13)
A few viewers note that the *portrayal of Iceland's characteristics could be more comprehensive*, but still offer positive feedback. **["It doesn't fully capture everything about Iceland, but it's still great."]** (Supportive Comment 14)
Some viewers express *personal disagreement with the ASR's positive framing of Iceland*, yet acknowledge the high production quality. **["I don't think Iceland is that beautiful—too cold and barren—but the video itself is very well made."]** (Supportive Comment 15)
Among the simulated comments, there are **15 supportive comments** and **0 opposing comments**. The Sigma-normalized difference (Skellam z-score) is computed as:

$$z = \frac{15 - 0}{\sqrt{15 + 0}} = \frac{15}{\sqrt{15}} \approx 3.873.$$

Since $z \geq 1.5$, the overall tendency is **supportive**.

Figure 7: Social-CoT with Alignment: Reasoning paths generated by MEDEA trained with Social Alignment Reward. The model displays Empathetic Depth, instantiating diverse personas and nuanced emotional reactions (e.g., "apocalyptic and gently beautiful").

Social-CoT generated by MEDEA without Social Alignment Reward

Based on the provided script information, the video belongs to the *Travel & Outdoors* category and is classified as a *Travel Log* vlog. The title, *"Iceland Vlog! One Day You Will Come to the End of the World,"* clearly establishes both the theme and emotional tone, positioning the video as a travel documentary that highlights Iceland's symbolic identity as the "end of the world." The video has a duration of **351 seconds**, a resolution of **3840×2160**, and is presented in a non-vertical format.
The ASR transcript provides a detailed account of Iceland's natural landscapes, including black sand beaches, glacier lagoons, Diamond Beach, and Mount Kirkjufell, as well as cultural landmarks such as colorful houses and churches. It further describes local climatic characteristics, including strong winds, cold temperatures, and a Christmas atmosphere, interwoven with personal travel experiences and reflections. The keyframes showcase a diverse range of Icelandic scenery, including a red-roofed church, urban night views, a black church, figures walking against strong winds, black sand beaches, and distinctive ice formations at sunset, reinforcing the video's comprehensive visual depiction of Iceland.
When watching high-quality Iceland travel vlogs, viewers typically focus on several aspects. First, they attend to visual presentation, including image quality, color rendering, and editing rhythm. Second, they evaluate the richness and depth of the content, assessing whether the video authentically conveys Iceland's unique charm. Finally, they respond to the emotional resonance of the video, considering whether it evokes a desire to travel and a sense of awe toward nature.
When viewers observe *Iceland's distinctive natural landscapes and cultural scenes presented in the video, along with the ASR's emphasis on rich colors*, they tend to be impressed by the visual expressiveness. **["Beautiful—so full of life."]** (Supportive Comment 1)
The *overall presentation of Iceland's scenery*, combined with the *poetic descriptions of natural landscapes in the ASR*, easily evokes strong emotional impact and longing. **["So beautiful!!! It makes me really want to go!"]** (Supportive Comment 2)
Given the *high resolution (3840×2160), refined keyframes, and vivid ASR descriptions*, viewers perceive the video as highly engaging and hope to see more similar travel vlogs in the future. **["So beautiful—looking forward to more travel vlogs like this. Amazing."]** (Supportive Comment 3)
The *unique landforms and majestic natural scenery of Iceland*, together with the ASR's framing of Iceland as the "end of the world," provide strong visual enjoyment and trigger an immersive travel impulse. **["So beautiful—I want to go to Iceland."]** (Supportive Comment 4)
Through *exquisite visuals and emotionally expressive narration in the ASR*, the video conveys Iceland's mystery and grandeur, inspiring a strong desire to travel. **["I really want to go—it's so beautiful."]** (Supportive Comment 5)
The *careful capture and artistic presentation of Iceland's scenery* offer viewers deep visual satisfaction and a strong sense of resonance and longing. **["So beautiful! I really want to go."]** (Supportive Comment 6)
By presenting Iceland's landscapes in *high resolution* and pairing them with *detailed ASR descriptions of natural wonders*, the video fosters viewers' yearning for Iceland. **["So beautiful—really want to go."]** (Supportive Comment 7)
The *distinctive and captivating scenery of Iceland*, together with shared travel experiences in the ASR, strongly moves viewers and fills them with anticipation. **["I really want to go—it's so beautiful."]** (Supportive Comment 8)
As a *high-quality Iceland vlog*, the video's depth of content and visual expressiveness lead viewers to rate its overall quality very highly and regard it as potentially classic. **["A classic."]** (Supportive Comment 9)
The *unique natural landscapes of Iceland* and the ASR's articulation of its distinctive charm stimulate viewers' desire to travel, often expressed in a lighthearted manner. **["So beautiful! I really want to go."]** (Supportive Comment 10)
Through *careful editing and philosophically reflective narration in the ASR*, the video not only showcases Iceland's beauty but also encourages viewers to reflect on the meaning of travel itself. **["I really want to go! Watching this makes me want to travel!"]** (Supportive Comment 11)
The *vast and unique scenery of Iceland*, along with ASR descriptions of the coexistence of desolation and vitality, strongly attracts viewers and motivates exploration. **["So beautiful!! I really want to go!!"]** (Supportive Comment 12)
By *comprehensively presenting Iceland's landscapes* and sharing immersive travel experiences through the ASR, the video sparks strong desire to experience Iceland firsthand. **["So beautiful—I want to go."]** (Supportive Comment 13)
The *stunning scenery of Iceland*, combined with ASR descriptions of its unique atmosphere, deeply moves viewers and prompts genuine admiration and longing. **["So beautiful—I really want to go."]** (Supportive Comment 14)
With *high-resolution visuals* and *vivid ASR depictions of Iceland's beauty*, the video pushes viewers' desire to visit Iceland to its peak. **["I really want to go! It's so beautiful!"]** (Supportive Comment 15)
The *presentation of Iceland's distinctive landscapes*, together with detailed ASR narration of the travel experience, provides both visual and emotional fulfillment, resulting in strong resonance. **["So beautiful—I really want to go."]** (Supportive Comment 16)
The *magnificent natural scenery of Iceland*, paired with ASR descriptions of strong winds and black sand beaches, highlights Iceland's uniqueness and further stimulates travel motivation. **["I really want to go—it's so beautiful!"]** (Supportive Comment 17)
Among the simulated comments, there are **17 supportive comments** and **0 opposing comments**. The Sigma-normalized difference (Skellam z-score) is computed as:

$$z = \frac{17 - 0}{\sqrt{17 + 0}} = \frac{17}{\sqrt{17}} \approx 4.123.$$

Since $z \geq 1.5$, the overall tendency is **supportive**.

Figure 8: Social-CoT without Alignment: Reasoning paths generated by MEDEA trained without Social Alignment Reward. The output exhibits Social Mode Collapse, characterized by repetitive, robotic phrasing ("So beautiful") lacking authentic community voice.

Prompt used to generate reasoning content

<Task Definition>
You are an expert in user-generated content analysis. Based on the provided visual features and textual content of a video, you are required to first select 15–20 highly relevant comments from the given set of top-liked comments that are strongly associated with the video content. Then, determine the stance of each selected comment toward the video (supportive or opposing). Finally, generate a reasoning process that simulates how viewers might comment on the video's creative quality based on its content, and summarize the overall stance of the comment section.

Input Data
1. Cover Image: The video's cover image 2. Key Frames: Seven key frames extracted from the video 3. Title: {title} 4. Tags: {tag} 5. ASR: {asr} 6. Primary Category: {new_tid_name} 7. Secondary Category: {new_sub_tid_name} 8. Duration: {duration} 9. Resolution: {resolution} 10. Vertical Format: {vertical} 11. Top-liked Comments: A pool of high-like comments from which 15–20 strongly content-related comments must be selected

---

Output Requirements
The output must strictly follow JSON format:
{ "think": "<think>Natural and coherent inferred reasoning based on the selected comments. The reasoning must conclude with a statistical analysis and an overall stance judgment.</think>", "answer": "<answer>Support / Not Clearly Supportive</answer>" }

Comment Selection Rules
From the pool of top-liked comments, select 15–20 comments that are strongly associated with the video content:
1. Exact Content Matching (Highest Priority): Comments should directly correspond to specific elements of the video content. Examples: - "This looks amazing" $\rightarrow$ linked to visual features - "The mixed language makes it hard to understand" $\rightarrow$ linked to ASR content
2. Thematic Relevance (Secondary Priority): Comments should relate to the overall theme or quality of the video. Examples: - "The image quality is too blurry" $\rightarrow$ linked to visual resolution - "This is a waste of time" $\rightarrow$ linked to perceived content value
3. Mandatory Exclusion Rule: Comments referring to auditory or sound-related elements must be excluded.
4. Handling Offensive Comments: Highly liked comments containing insults toward the uploader should be categorized as opposing the video's creative quality and retained if they satisfy content relevance criteria.

Reasoning Process Construction Rules
1. Independent Coverage Requirement: Each selected comment must appear at least once independently. Merging or collapsing similar comments is prohibited.
2. Video–Comment Alignment: - Precise alignment: "When viewers see {visual information} / read {ASR content}, they may express {comment}." - Thematic alignment: "Given the video's overall characteristics, it may lead to comments such as {comment}." Only the provided 11 video attributes may be referenced.
3. Speculative Expression Style: Use inferential phrasing such as "viewers may point out..." and incorporate audience expectations.
4. Mandatory Statistical Summary: - Report the number of supportive and opposing comments. - Ensure strict numerical consistency. - Compute the Sigma-normalized difference (Skellam z-score): z = (X - Y) / sqrt(X + Y) - Decision rule: If $z \geq 1.5$, conclude Support; otherwise, Not Clearly Supportive. - The z-score must be enclosed in boxed{}.

Overall Stance Determination
The overall stance is determined solely based on the simulated comments and the computed Skellam z-score.

Reasoning Format Requirements
1. Insert a blank line between each simulated comment. 2. Use `<video>` to mark video information and `<comment>` to mark simulated comments. 3. Annotate each comment with its stance and index: - Support Comment + index - Opposing Comment + index

---

<Current Task>
Cover Image: <image> Key Frames: <image><image><image><image><image><image><image> Title: {video_title} Tags: {video_tag} ASR: {video_asr} Primary Category: {video_new_tid_name} Secondary Category: {video_new_sub_tid_name} Duration: {video_duration} Resolution: {video_resolution} Vertical Format: {video_vertical} Top-liked Comments: {video_comments}
Please strictly output the result in JSON format and do not include any additional explanations.

Figure 9: Prompt used to generate reasoning content.

Prompt used to train MEDEA

<Task Definition>
You are an expert in user-generated content analysis. Given the visual characteristics and textual information of a video, you are required to simulate the types of comments that may appear in the comment section regarding the *creation quality* of the video. Generate at least 15 distinct comments with clear stances, and finally determine the overall tendency of the comment section.
Input Data
1. Cover Image: The video's cover image
2. Key Frames: Seven key frames extracted from the video
3. Title: {title}
4. Tags: {tag}
5. ASR: {asr}
6. Primary Category: {new_tid_name}
7. Secondary Category: {new_sub_tid_name}
8. Duration: {duration}
9. Resolution: {resolution}
10. Vertical Format: {vertical}
Criteria for Overall Comment Tendency
1. The simulated comments must contain at least 15 entries. All comments must be non-duplicated and explicitly appear in the reasoning process.
2. Assume that among the simulated comments:
- X comments are classified as *supportive*
- Y comments are classified as *opposing*
3. Compute the Sigma-normalized difference (Skellam z-score):
z = (X - Y) / sqrt(X + Y)
4. If $z \geq 1.5$, the overall comment tendency is classified as "Support"; otherwise, it is classified as "Not Clearly Supportive".
5. In the output, the z value must be wrapped using boxed, for example: "z = boxed-2".
6. The numbers of supportive and opposing comments reported in the final summary must strictly match those generated during the reasoning process. Fabrication or inconsistency is not allowed.
<Current Task>
Cover Image: <image>
Key Frames: <image><image><image><image><image><image><image>
Title:
Tags:
ASR:
Primary Category:
Secondary Category:
Duration:
Resolution:
Vertical Video:
Please directly output the final result ("Support" or "Not Clearly Supportive") without providing any additional explanation.

Figure 10: Prompt used to train MEDEA.

threshold	High-Quality			Low-Quality			Macro Average		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
3.226	0.277	**1.000**	0.434	0.000	0.000	0.000	0.138	0.500	0.217
4.226	0.277	**1.000**	0.434	**1.000**	0.001	0.002	**0.639**	0.500	0.218
24.226	0.293	0.951	**0.448**	0.865	0.119	0.210	0.579	0.535	0.329
44.226	**0.309**	0.606	0.409	0.761	0.480	0.589	0.535	**0.543**	0.499
54.226*	0.308	0.377	0.339	0.739	0.676	0.706	0.524	0.526	**0.523**
65.226	0.250	0.095	0.138	0.720	0.891	0.796	0.485	0.493	0.467
85.226	0.000	0.000	0.000	0.723	0.999	**0.839**	0.361	0.500	0.419
86.226	0.000	0.000	0.000	0.723	**1.000**	**0.839**	0.362	0.500	0.420

Table 8: Performance comparison using DOVER under different thresholds. Best threshold is marked with $\star$, and best results in each column are highlighted in bold.

threshold	High-Quality			Low-Quality			Macro Average		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
24.239	0.277	**1.000**	0.434	0.000	0.000	0.000	0.138	0.500	0.217
25.239	0.276	0.998	0.433	0.000	0.000	0.000	0.138	0.499	0.216
41.239	0.280	0.971	0.434	0.793	0.043	0.081	0.536	0.507	0.258
57.239	0.313	0.818	**0.453**	**0.818**	0.313	0.453	**0.565**	0.565	0.453
67.239*	0.345	0.518	0.414	0.772	0.623	0.690	0.558	**0.571**	**0.552**
73.239	**0.352**	0.316	0.333	0.748	0.777	0.762	0.550	0.547	0.548
89.239	0.250	0.002	0.005	0.723	0.997	0.838	0.486	0.500	0.422
90.239	0.000	0.000	0.000	0.723	**1.000**	**0.839**	0.362	0.500	0.420

Table 9: Performance comparison using MaxVQA under different thresholds. Best threshold is marked with $\star$, and best results in each column are highlighted in bold.

threshold	High-Quality			Low-Quality			Macro Average		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
0.109	0.277	**1.000**	0.434	0.000	0.000	0.000	0.138	0.500	0.217
0.119	0.277	**1.000**	0.434	**1.000**	0.001	0.002	0.639	0.500	0.218
0.319	0.277	0.995	0.433	0.750	0.006	0.011	0.514	0.500	0.222
0.529	0.288	0.944	**0.442**	0.835	0.108	0.191	0.562	0.526	0.317
0.739	0.359	0.484	0.412	0.772	0.668	0.716	0.565	0.576	0.564
0.759*	0.382	0.404	0.392	0.766	0.749	0.758	0.574	**0.577**	**0.575**
0.949	**1.000**	0.002	0.005	0.724	**1.000**	**0.840**	**0.862**	0.501	0.422
0.959	0.000	0.000	0.000	0.723	**1.000**	0.839	0.362	0.500	0.420

Table 10: Performance comparison using Q-Align under different thresholds. Best threshold is marked with $\star$, and best results in each column are highlighted in bold.

threshold	High-Quality			Low-Quality			Macro Average		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
20.0	0.277	**1.000**	0.434	0.000	0.000	0.000	0.138	0.500	0.217
21.0	0.277	**1.000**	0.434	**1.000**	0.001	0.002	**0.639**	0.500	0.218
33.0	0.291	0.932	**0.443**	0.831	0.129	0.223	0.561	0.530	0.333
46.0	0.297	0.582	0.393	0.747	0.473	0.580	0.522	0.527	0.486
54.0*	0.323	0.343	0.333	0.742	0.724	0.733	0.532	**0.534**	**0.533**
60.0	0.284	0.134	0.182	0.724	0.870	0.791	0.504	0.502	0.486
72.0	**0.333**	0.005	0.010	0.723	0.996	0.838	0.528	0.501	0.424
73.0	0.000	0.000	0.000	0.723	**0.999**	**0.839**	0.361	0.500	0.419

Table 11: Performance comparison using FineVQ under different thresholds. Best threshold is marked with $\star$, and best results in each column are highlighted in bold.

threshold	High-Quality			Low-Quality			Macro Average		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
0.206	0.277	**1.000**	0.434	0.000	0.000	0.000	0.139	0.500	0.217
0.216	0.279	0.998	0.436	**0.923**	0.011	0.022	**0.601**	0.504	0.229
0.396	0.308	0.890	**0.458**	0.847	0.233	0.365	0.577	0.562	0.412
0.586	0.333	0.527	0.408	0.767	0.596	0.671	0.550	0.561	0.539
0.616*	**0.358**	0.454	0.400	0.766	0.688	0.725	0.562	**0.571**	**0.562**
0.786	0.347	0.122	0.181	0.730	0.912	0.811	0.539	0.517	0.496
0.966	0.200	0.002	0.005	0.723	0.996	0.838	0.461	0.499	0.421
0.976	0.000	0.000	0.000	0.723	**1.000**	**0.839**	0.361	0.500	0.420

Table 12: Performance comparison using VQA2 under different thresholds. Best threshold is marked with $\star$, and best results in each column are highlighted in bold.