ADALIGN-BENCH: A BENCHMARK FOR HUMAN-ALIGNED ADVERTISEMENT SUBJECTIVE VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Subjective video judgments reflect viewers' experiences and perceptions, raising the question of whether multimodal large language models (MLLMs) align with human evaluations. Existing subjective benchmarks commonly use aggregated human ratings, which measure population-level alignment but offer limited insight into viewer differences and conditions of model–human divergence. We introduce **AdAlign**, a benchmark centered on population-level subjective rating alignment, using individual responses, psychological traits, and information-processing-related ratings to examine variation in model–human correspondence that aggregate scores may obscure. **AdAlign** comprises annotations by 225 participants on 393 advertisement videos, covering creativity and multimodal perceptual quality alongside open-ended impressions and self-reported viewing experiences. Current MLLMs align more closely with human ratings of creativity than of perceptual quality, capturing broad rating trends while exhibiting discrepancies in rating agreement and absolute scores. Alignment declines in both higher-rated and higher-disagreement video subsets, more markedly in the former. Under generic viewer prompts, model ratings also exhibit trait-related patterns of correspondence with human judgments. An exploratory case study suggests that correspondence with human information-processing reports may support creativity alignment, while illustrating differences between human and model response-level correlation structures. Together, these findings highlight variation in subjective rating alignment that aggregate performance alone may obscure.

1 INTRODUCTION

Multimodal large language models (MLLMs) have rapidly advanced video understanding, with benchmark evaluation expanding from factual recognition to complex reasoning. Existing benchmarks primarily assess understanding of verifiable video content (Caba Heilbron et al., 2015; Kay et al., 2017; Liang et al., 2025; Liu et al., 2025; Krishna et al., 2017; Li et al., 2025; Xu et al., 2016; Wang et al., 2019; Yu et al., 2019; Mangalam et al., 2023; Li et al., 2024b; Wu et al., 2025; Fu et al., 2025), laying the foundation for objective video understanding. Recent work extends to higher-level cognition, including implicit information understanding (Long et al., 2025), causal-temporal narrative modeling (Nadeem et al., 2024), and narrative understanding (Zhang et al., 2025). Yet subjective video understanding, where judgments are grounded in viewers' experiences and perceptions, remains comparatively underexplored.

Unlike objective tasks, where inter-annotator disagreement is often treated as noise around a correct answer, subjective disagreement can reflect meaningful variation in perception and evaluation. Existing subjective benchmarks, such as those for aesthetic judgment (Sun et al., 2025; Xie et al., 2025; Jiang & Chen, 2025; An et al., 2026), commonly evaluate MLLM–human alignment against aggregated ratings. These evaluations measure population-level alignment but offer limited insight into viewer differences and the conditions under which model judgments diverge from human evaluations. We therefore evaluate population-level subjective rating alignment while using individual responses, psychological traits, and information-processing-related ratings to diagnose heterogeneity and failure patterns beyond aggregate performance. Specifically, we examine alignment across video subsets differing in human disagreement and mean ratings, trait-related variation in model–

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

human correspondence without supplying traits to models, and similarities and differences in the relationships among human and model information-processing-related ratings.

To address these questions, we introduce **AdAlign**, a benchmark for MLLM–human alignment in subjective video understanding. Short-form advertisements provide a compact yet demanding setting, conveying information, attracting attention, evoking emotion, and communicating symbolic meaning within a limited duration. Their reliance on visual rhetoric, symbolic meaning transfer, and the integration of information and emotion (McCracken, 1986; Scott, 1994; Teixeira et al., 2014; Tellis et al., 2019) makes them a useful testbed for human-centered subjective evaluation. AdAlign comprises annotations by 225 participants on 393 advertisement videos across two core dimensions: creativity and multimodal perceptual quality. It includes individual ratings, psychological trait measures, open-ended impressions, and self-reported information-processing experiences. These annotations support population-level evaluation alongside analyses of viewer differences, salient impressions, and relationships among information-processing-related responses.

Current MLLMs align more closely with human ratings of creativity than of multimodal perceptual quality. They capture broad rating trends but exhibit discrepancies in rating agreement and absolute scores. As evaluation is restricted to higher-rated or higher-disagreement videos, alignment declines in both cases, more markedly in the former. Even under generic viewer prompts, model ratings exhibit trait-related patterns of correspondence with human judgments. An exploratory case study further suggests that correspondence with human information-processing reports may support creativity evaluation, while revealing differences between human and model response-level correlation structures.

Our main contributions are as follows:

(i) **A benchmark for subjective video evaluation.** We introduce **AdAlign**, a human-annotated advertisement video benchmark covering **creativity** and **multimodal perceptual quality**. It supports population-level rating evaluation while retaining individual responses and psychological traits to examine variation in human judgments and model–human correspondence.
(ii) **A multi-level analysis of subjective rating alignment.** We examine alignment across video subsets differing in human disagreement and mean ratings, and analyze trait-related variation in model–human correspondence, revealing differences that aggregate performance alone may obscure.
(iii) **Evaluation beyond content ratings.** We complement rating-based evaluation with salient-impression extraction and information-processing-related analyses, examining coverage of human-reported impressions and similarities and differences in response-level correlation structures.

2 Related Work

Video understanding benchmarks. Existing video understanding benchmarks have primarily focused on objective and verifiable video content. Representative tasks include recognition-oriented benchmarks such as action and event understanding (Caba Heilbron et al., 2015; Kay et al., 2017; Liang et al., 2025; Liu et al., 2025), localization-oriented tasks such as temporal grounding (Li et al., 2025), language-centered tasks such as video captioning and question answering (Xu et al., 2016; Krishna et al., 2017; Wang et al., 2019; Yu et al., 2019; Mangalam et al., 2023), as well as broader benchmarks for holistic video understanding (Li et al., 2024b; Wu et al., 2025; Fu et al., 2025). These benchmarks establish a foundation for evaluating video content understanding. More recent work has extended video understanding toward higher-level reasoning, including implicit information understanding (Long et al., 2025), causal-temporal narrative modeling (Nadeem et al., 2024), and narrative understanding (Zhang et al., 2025), moving beyond surface-level semantic recognition. Related work on subjective visual and multimodal evaluation has also examined aesthetics and perceptual quality (Sun et al., 2025; Xie et al., 2025; Jiang & Chen, 2025; An et al., 2026). Such evaluations provide useful measures of model correspondence with human judgments, but aggregate performance alone offers limited insight into variation across viewers. Research on annotation disagreement further motivates retaining diverse judgments as informative variation rather than treating them solely as noise (Pavlick & Kwiatkowski, 2019; Uma et al., 2021). AdAlign complements these efforts by combining population-level subjective rating evaluation with analyses of individual responses and viewer characteristics in advertisement videos.

Subjective video content evaluation: Creativity and multimodal perceptual quality. Subjective video evaluation depends not only on what a video conveys, but also on how it is expressed and perceived by viewers. This is particularly relevant to short-form advertisements, which combine creative design with multimodal presentation. Prior marketing research commonly characterizes advertising creativity in terms of *divergence* and *relevance* (Smith et al., 2007; Yang & Smith, 2009). We focus on *divergence*—including originality, flexibility, synthesis, elaboration, and artistic value—to evaluate viewers' perceptions of the video's creative expression. In parallel, research on consumer responses highlights the role of visual presentation (Meyers-Levy & Peracchio, 1992; Elder & Krishna, 2012; Min et al., 2025; Sevilla & Meyer, 2020; Chu et al., 2021; Jia et al., 2020; Roggeveen et al., 2015), visual aesthetics (Li & Xie, 2020; Hagen, 2021), music (MacInnis & Park, 1991; Huang & Labroo, 2020), and sensory language (Cascio Rizzo et al., 2023). These findings motivate attention to the perceptual aspects of audiovisual presentation alongside creative expression. Together, **creativity** and **multimodal perceptual quality** provide two complementary perspectives on subjective advertisement evaluation and a basis for studying MLLM–human alignment.

Information processing and individual differences in video understanding. Recent user-centered studies of video have examined induced sentiment (Jia et al., 2024), engagement and popularity (Li et al., 2024a; Xu et al., 2025b), and potential audiences (Long et al., 2025). These tasks characterize audience responses or related attributes, motivating a complementary examination of viewers' experiences during evaluation. Research on information processing and persuasion provides relevant perspectives on curiosity and attention (Loewenstein, 1994), affective responses (Yang & Smith, 2009; Tellis et al., 2019), closure-related processing (Kardes et al., 2004; Yang & Smith, 2009), and narrative transportation and processing (Green & Brock, 2000; Escalas, 2004). Individual differences, including personality and motivational orientations, can also shape responses to persuasive messages (Aaker & Lee, 2001; Hirsh et al., 2012; Matz et al., 2017). Together, these findings motivate complementing population-level rating evaluation with viewers' self-reported information-processing experiences and psychological trait measures. In AdAlign, these annotations support analyses of how model–human rating correspondence varies with viewer traits and how the relationships among information-processing-related ratings compare between models and humans.

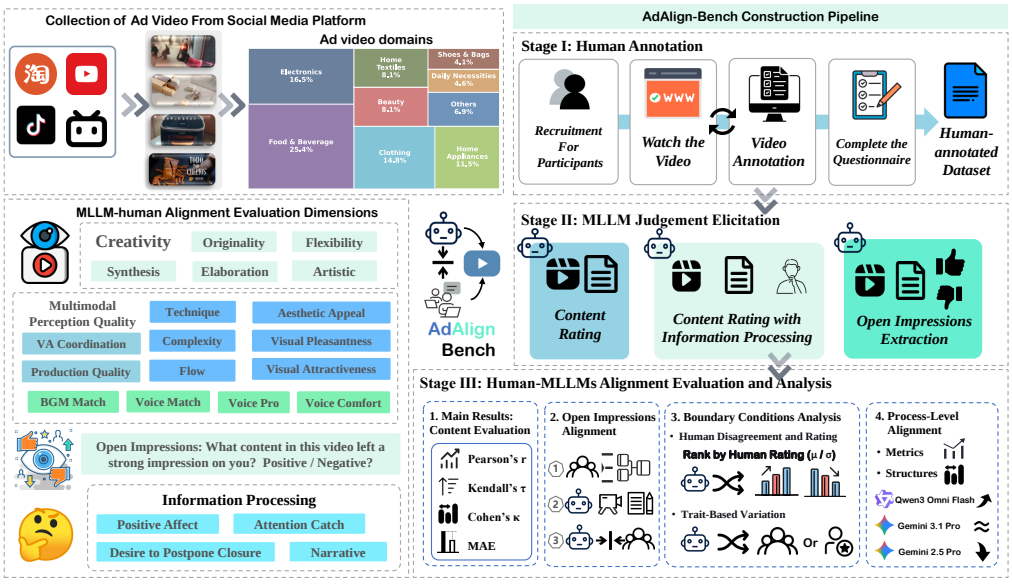


Figure 1: Overview of **AdAlign**. AdAlign evaluates MLLM–human alignment through content ratings, open impressions, and information-processing, centered on **Creativity** and **Multimodal Perceptual Quality**. The benchmark consists: Human Annotation, MLLM Judgement Elicitation, and Alignment Evaluation and Analysis.

Table 1: Dimensions of Content Evaluation, Creativity and Multimodal Perceptual Quality.

Dimension	Definition
Creativity (Smith et al., 2007; Yang & Smith, 2009)	
Originality	The ad's ideas are rare, surprising, or move away from the obvious and commonplace.
Flexibility	The ad had different ideas and shifted from one type of subject matter to another.
Synthesis	The ad combined or connected normally unrelated objects or ideas.
Elaboration	The ad provided numerous details. The ad finished, extended, and detailed basic ideas so they become more intricate or sophisticated.
Artistic Value	The ad had striking visual and/or verbal elements.
Multimodal Perceptual Quality (Visual)	
Technique	The visuals demonstrate professional production techniques, such as framing, composition, camera work, or shot design (Meyers-Levy & Peracchio, 1992; Elder & Krishna, 2012; Sevilla & Meyer, 2020; Chu et al., 2021).
Complexity	The visuals contain rich compositional elements, layered scenes, or a high density of visual information (Min et al., 2025).
Flow	The visuals progress smoothly, with coherent editing, movement, and pacing (Roggeveen et al., 2015; Jia et al., 2020).
Aesthetic	Nuanced visual aesthetic perception deconstructed into: **Aesthetic Appeal**, **Visual Pleasantness**, and **Visual Attractiveness** (Li & Xie, 2020; Hagen, 2021).
Multimodal Perceptual Quality (Audio) (MacInnis & Park, 1991; Huang & Labroo, 2020; Cascio Rizzo et al., 2023)	
BGM-Visual match	The background music matches the visual content, mood, and rhythm of the video.
Speech comfort	The narration or spoken voice is comfortable to listen to.
Speech fit	The narration or spoken voice matches the product/ad well.
Speech professionalism	The narration or spoken explanation sounds professional.

3 ADALIGN-BENCH

3.1 DUAL-LAYER MULTIDIMENSIONAL SUBJECTIVE EVALUATION

To systematically evaluate subjective understanding of short-form video advertisements, we develop a dual-layer multidimensional evaluation framework grounded in research on advertising creativity, media perception, persuasion, and information processing. Viewers form evaluations of advertisement content and presentation while also experiencing cognitive and affective responses during viewing. Accordingly, we organize the evaluation into two layers: content evaluation and information-processing evaluation. Content ratings serve as the primary task for assessing population-level alignment, while information-processing-related ratings support complementary analyses of response structures and their relationship to content evaluations.

Content evaluation measures viewers' direct subjective evaluations of advertisement content and presentation across two core dimensions: creativity and multimodal perceptual quality. **Information-processing evaluation** characterizes viewers' self-reported experiences during advertisement viewing across four dimensions: attention catch, positive affect, narrative, and desire to postpone closure (DPC). Detailed definitions are provided in Tables 1, and the annotation items are listed in Appendix B.1.

3.2 DATASET CONSTRUCTION

Video Collection. To support subjective evaluation of short-form video advertisements, we collected videos with consideration of creative quality, production quality, and public accessibility. Videos were sourced from major social media platforms, including Douyin, Taobao, Bilibili, and YouTube. For each video, we collected the video and its associated metadata. To facilitate comprehension, all selected videos were provided with subtitles in the participants' native language. The final collection comprises 393 videos.

Human Annotation Process. We collected subjective evaluations from ordinary viewers in controlled laboratory sessions to capture their evaluations and viewing experiences. Each session consisted of videos evaluation followed by questionnaire completion. Participants provided content ratings, information-processing-related ratings, and open-ended impressions through a dedicated interface (Appendix B). To reduce framing effects, participants saw the annotation items without the corresponding dimension names or formal definitions. To reduce priming, additional product information was shown only after the video ended, and replay was disabled on the response page to elicit immediate post-viewing impressions.

Participants then completed questionnaires measuring individual differences relevant to subjective judgment, including Big Five personality traits, regulatory focus, consumer decision-making styles, and emotional contagion. These measures were used to analyze how model–human rating correspondence varies with viewer traits and were not provided as model inputs. The study received ethics approval and was conducted with informed consent. Each session lasted approximately 80 minutes, and participants were compensated at approximately $8 per hour. Data were collected from January 10 to January 20, 2026. In total, 225 participants provided subjective evaluations of 393 videos, with an average of approximately 13 participants per video.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Evaluated Models. We evaluate a range of multimodal large language models, including **GPT-5.4** (Singh et al., 2025), **Kimi-K2.5** (Team et al., 2026), the **Gemini** family (Comanici et al., 2025; Deepmind, 2025; 2026), **Qwen3-VL** (Bai et al., 2025), **Qwen3.5** (Team, 2026b), and **Qwen-Omni** (Xu et al., 2025a; Team, 2026a). Frames are sampled at 2 fps (1 fps for Gemini); Gemini and Qwen-Omni models additionally receive audio input.

Rating-Based Evaluation. We evaluate population-level alignment against **mean opinion scores (MOS)**, obtained by averaging human ratings for each video and dimension, and report by dimension. Our primary metric is **Kendall's** τ, which measures agreement with the MOS-based video ranking. We also report **Pearson correlation** for linear association, **linear-weighted Cohen's** κ for chance-corrected rating agreement, and **mean absolute error (MAE)** for absolute score differences. For each rating prompt, we average three model responses before evaluation. All models use temperature 1, top-p 0.9, and enabled thinking mode. Metric definitions are in Appendix E.2.

Salient-Impression Evaluation. We assess how well MLLMs recover salient impressions expressed in human responses under a fixed output budget. The evaluation proceeds in three stages. First, an evaluator organizes the 12–34 raw human responses for each video into semantically distinct reference impression points. Second, each evaluated model generates exactly five ranked impression points for the same video and is instructed to make them as distinct as possible. Third, an evaluator independently matches each model-generated point to its best-matching reference point and assigns a score of 1, 0.5, or 0 for a strict, relaxed, or no match, respectively. We implement this procedure using three evaluator pipelines—GPT-5.4/GPT-5.4-mini, Gemini-3.5-Flash, and DeepSeek-V4-Pro—each constructing its own reference set and performing semantic matching. Because outputs are matched independently, multiple points may correspond to the same reference impression. We credit only the earliest match to each reference point; later outputs matching that point retain their positions in the fixed output budget but receive no additional credit. This separates semantic matching from scoring and penalizes repeated coverage of the same human impression. We report Recall@K and Precision@K for $K \in \{3, 5\}$ separately for each pipeline. These metrics measure correspondence with evaluator-organized reference impressions under each pipeline's matching procedure. Support-weighted results are provided in Table 11.

Prompt Settings. We use five settings: creativity and multimodal perceptual quality evaluation, each with and without information-processing items, plus salient-impression extraction. The two content dimensions use separate prompts to limit cross-task interference. In the expanded prompts, original content ratings and added information-processing ratings are analyzed separately. All rating prompts adopt an ordinary-viewer perspective without individual trait information. Full prompts are provided in Appendix I.

4.2 MAIN RESULTS

Variation across models. Table 2 shows distinct performance profiles across dimensions. The **Gemini** family achieves the highest Kendall's τ on many dimensions. **GPT-5.4** leads on synthesis, artistic value, and aesthetic appeal, while **Qwen3.5-Plus** is competitive across several creativity and perceptual-quality dimensions. **Qwen3.5-Omni-Plus** shows a markedly uneven profile across dimensions: strongest on complexity, but substantially lower on flow than on any other visual dimension.

Table 2: Performance of models on Creativity and Multimodal Perceptual Quality metrics. Best results are in bold; second-best results are underlined. The same convention is used throughout.

Metric	Model	Creativity						Multimodal Perceptual Quality											
		Divergence					Generally	Visual						Audio-Visual Match				Generally	
		Orig.	Flex.	Synth.	Elab.	Art.		Cmplx	Tech	Flow	Appeal	Plea	Attr	BGM	SMatch	SComf	SPro	VACoord	ProdQ
Kendall's τ	**Closed-source Models**																		
	Gemini-2.5-Pro	0.641	0.637	0.613	0.467	0.499	0.675	0.560	**0.604**	0.424	0.450	0.406	0.443	0.378	0.200	0.462	0.390	0.398	**0.604**
	Gemini-3-Flash	**0.676**	0.597	0.636	0.388	0.439	**0.681**	0.541	0.587	0.393	0.462	0.459	0.489	0.466	0.371	0.435	**0.402**	0.448	0.584
	Gemini-3.1-Pro	0.648	**0.642**	0.639	**0.549**	0.532	0.667	0.548	0.596	**0.527**	0.526	**0.490**	**0.518**	**0.485**	**0.399**	**0.549**	0.380	**0.518**	0.552
	GPT-5.4	0.636	0.636	**0.645**	0.423	**0.551**	0.635	0.594	0.571	0.385	**0.563**	0.453	0.486	0.394	0.279	0.270	0.386	0.461	0.493
	Qwen3.5-Omni-Plus	0.613	0.424	0.491	0.327	0.298	0.604	**0.629**	0.352	0.029	0.308	0.279	0.327	0.304	0.067	0.132	0.233	0.198	0.269
	Qwen3.5-Plus	0.607	0.573	0.612	0.423	0.467	0.629	0.533	0.519	0.439	0.435	0.420	0.478	0.445	0.373	0.365	0.325	0.457	0.456
	Open-source Models																		
	Kimi-k2.5	0.617	0.595	0.618	0.414	0.455	0.630	0.493	0.516	0.415	0.390	0.362	0.416	0.358	0.219	0.296	0.202	0.462	0.475
	Qwen3-VL-30B	0.473	0.483	0.530	0.336	0.372	0.490	0.435	0.436	0.333	0.383	0.407	0.442	0.252	0.251	0.286	0.249	0.376	0.428
	Qwen3-VL-32B	0.534	0.543	0.583	0.339	0.469	0.570	0.542	0.539	0.341	0.468	0.439	0.481	0.270	0.276	0.256	0.304	0.404	0.535
	Qwen3-VL-235B	0.483	0.472	0.533	0.387	0.383	0.482	0.458	0.464	0.359	0.421	0.414	0.472	0.249	0.192	0.203	0.200	0.367	0.453
	Qwen3-Omni-Flash	0.518	0.485	0.500	0.388	0.452	0.538	0.448	0.418	0.328	0.402	0.344	0.385	0.367	0.195	0.218	0.208	0.294	0.345
	Qwen3.5-27B	0.484	0.463	0.554	0.307	0.410	0.520	0.404	0.364	0.188	0.297	0.226	0.333	0.215	0.253	0.252	0.175	0.294	0.366
	Qwen3.5-35B	0.536	0.503	0.585	0.284	0.378	0.581	0.357	0.412	0.225	0.329	0.329	0.399	0.237	0.233	0.127	0.191	0.348	0.375
Pear Corr.	**Closed-source Models**																		
	Gemini-2.5-Pro	0.824	**0.850**	0.810	0.655	0.695	**0.853**	0.731	0.796	0.656	0.632	0.588	0.615	0.533	0.343	0.619	0.561	0.422	0.723
	Gemini-3-Flash	**0.842**	0.815	0.825	0.611	0.638	0.845	0.691	0.782	0.623	0.642	0.637	0.675	0.625	0.500	0.546	**0.566**	0.626	**0.773**
	Gemini-3.1-Pro	0.810	0.840	**0.827**	**0.743**	**0.726**	0.842	0.700	**0.800**	**0.697**	**0.701**	**0.662**	**0.731**	**0.667**	**0.528**	**0.684**	0.539	**0.673**	0.754
	GPT-5.4	0.802	0.805	0.825	0.553	0.711	0.802	0.742	0.747	0.506	0.694	0.601	0.623	0.540	0.323	0.335	0.495	0.569	0.649
	Qwen3.5-Omni-Plus	0.743	0.646	0.690	0.492	0.472	0.738	**0.776**	0.610	0.079	0.445	0.426	0.467	0.464	0.077	0.223	0.353	0.318	0.442
	Qwen3.5-Plus	0.798	0.784	0.815	0.602	0.653	0.820	0.692	0.720	0.587	0.589	0.589	0.649	0.631	0.485	0.497	0.429	0.625	0.642
	Open-source Models																		
	Kimi-k2.5	0.791	0.797	0.811	0.574	0.641	0.804	0.616	0.662	0.534	0.536	0.502	0.563	0.513	0.288	0.328	0.308	0.579	0.614
	Qwen3-VL-30B	0.679	0.692	0.733	0.489	0.548	0.697	0.594	0.639	0.485	0.512	0.533	0.596	0.346	0.366	0.408	0.365	0.526	0.633
	Qwen3-VL-32B	0.726	0.742	0.784	0.462	0.643	0.761	0.730	0.773	0.547	0.611	0.581	0.666	0.382	0.412	0.440	0.439	0.586	0.756
	Qwen3-VL-235B	0.665	0.659	0.728	0.519	0.523	0.666	0.616	0.690	0.516	0.565	0.542	0.634	0.343	0.240	0.283	0.269	0.514	0.662
	Qwen3-Omni-Flash	0.708	0.701	0.688	0.528	0.643	0.722	0.598	0.593	0.448	0.535	0.471	0.562	0.519	0.255	0.283	0.297	0.434	0.490
	Qwen3.5-27B	0.675	0.673	0.756	0.440	0.568	0.711	0.542	0.578	0.283	0.418	0.380	0.441	0.331	0.340	0.394	0.271	0.450	0.502
	Qwen3.5-35B	0.710	0.694	0.782	0.371	0.510	0.766	0.551	0.600	0.283	0.484	0.483	0.555	0.327	0.335	0.261	0.303	0.508	0.533
Cohen's Kappa	**Closed-source Models**																		
	Gemini-2.5-Pro	0.503	0.542	0.440	0.397	0.390	0.491	**0.332**	0.335	0.127	0.292	0.248	0.069	0.053	0.075	0.319	0.212	0.021	0.174
	Gemini-3-Flash	**0.639**	0.547	0.541	0.318	0.384	0.578	0.185	0.143	0.056	0.208	0.134	0.064	0.032	0.017	0.131	0.037	0.012	0.075
	Gemini-3.1-Pro	0.541	**0.548**	0.524	**0.497**	**0.459**	0.579	0.289	**0.352**	**0.224**	**0.328**	**0.302**	**0.222**	0.153	0.141	**0.417**	**0.262**	0.102	**0.261**
	GPT-5.4	0.578	0.432	0.501	0.307	0.298	0.456	0.180	0.214	0.188	0.125	0.159	0.106	**0.313**	0.029	0.180	0.226	0.233	0.123
	Qwen3.5-Omni-Plus	0.315	0.309	0.248	0.278	0.185	0.236	0.078	0.105	0.011	0.083	0.087	0.017	0.020	0.008	0.042	0.170	0.005	0.031
	Qwen3.5-Plus	0.589	0.519	**0.558**	0.354	0.357	**0.588**	0.212	0.226	0.168	0.142	0.123	0.102	0.297	0.135	0.236	0.164	0.123	0.121
	Open-source Models																		
	Kimi-k2.5	0.482	0.424	0.523	0.230	0.302	0.410	0.125	0.098	0.068	0.077	0.076	0.032	0.117	0.057	0.074	0.107	0.075	0.068
	Qwen3-VL-30B	0.420	0.362	0.451	0.228	0.276	0.324	0.150	0.115	0.088	0.079	0.070	0.066	0.072	0.076	0.080	0.080	0.155	0.067
	Qwen3-VL-32B	0.477	0.487	0.545	0.243	0.406	0.522	0.297	0.262	0.195	0.086	0.073	0.075	0.173	0.130	0.183	0.228	**0.329**	0.154
	Qwen3-VL-235B	0.432	0.401	0.497	0.300	0.311	0.434	0.330	0.337	0.213	0.173	0.136	0.124	0.089	**0.146**	0.121	0.157	0.287	0.221
	Qwen3-Omni-Flash	0.300	0.180	0.288	0.095	0.220	0.198	0.046	0.042	0.043	0.026	0.028	0.022	0.012	0.016	0.040	0.079	0.019	0.018
	Qwen3.5-27B	0.424	0.359	0.527	0.240	0.314	0.418	0.072	0.118	0.064	0.079	0.099	0.024	0.071	0.089	0.105	0.122	0.083	0.068
	Qwen3.5-35B	0.478	0.457	0.542	0.194	0.250	0.515	0.089	0.179	0.067	0.129	0.131	0.058	0.137	0.122	0.066	0.209	0.213	0.134

Variation across dimensions. For **creativity**, models generally show stronger correspondence with human MOS on originality, flexibility, and synthesis than on elaboration and artistic value. This suggests that alignment is stronger when judging the novelty and combination of ideas than when judging how those ideas are developed and artistically expressed. For **multimodal perceptual quality**, technique generally yields higher ranking correspondence than flow, and visual pleasantness shows the weakest alignment among the aesthetic dimensions. Gemini models lead on several audio-related items, but the evaluated Qwen-Omni models do not consistently outperform vision–language models. Thus, among the evaluated systems, audio-input support is not accompanied by a consistent advantage on audio-related items.

Variation across metrics. Weaker alignment on perceptual quality than on creativity is broadly consistent across metrics, with the contrast particularly pronounced in linear-weighted κ. The metrics nevertheless reveal different aspects of alignment: Pearson's r measures linear association, Kendall's τ measures ranking correspondence, and weighted κ assesses rating-category agreement. Strong correlation therefore need not imply close agreement in score levels. For example, Gemini-3-Flash shows substantial ranking correspondence with human MOS on production quality, yet assigns higher scores on average. MAE and signed-bias results further reveal absolute score discrepancies and predominantly positive biases on perceptual-quality items (Appendix E.4). Together, these results show that models can rank videos similarly to human viewers while still assigning substantially different scores.

Salient-impression extraction. Table 3 reports results from three evaluator pipelines; @3 and @5 consider the first three and all five model outputs, respectively. The number and support-count distribution of consolidated reference impressions are reported in Appendix Table 15. **GPT-5.4** achieves the highest R@5 and P@5 under all three pipelines. At @3, GPT-5.4 leads on both metrics under the GPT pipeline, and **Qwen3.5-Omni-Plus** leads under the Gemini pipeline. Under the Gemini and DeepSeek pipelines, **Qwen3.5-Omni-Plus** has the highest R@3 and P@3. Thus, the full-output comparison identifies a consistent leader, whereas the top-three comparison is more sensitive to how reference impressions are organized and matched. The models leading on many MOS-based dimensions do not consistently lead in salient-impression extraction. These results distinguish two

Table 3: Results on open impression extraction measured by Recall@K and Precision@K.

Model	GPT-5.4				Gemini-3.5-Flash				DeepSeek-V4-Pro			
	R@3	P@3	R@5	P@5	R@3	P@3	R@5	P@5	R@3	P@3	R@5	P@5
Closed-source Models												
Gemini-2.5-Pro	0.1786	0.3880	0.2390	0.3127	0.1728	0.3643	0.2371	0.3038	0.1684	0.3013	0.2227	0.2554
Gemini-3-Flash	0.1786	0.3902	0.2362	0.3107	0.1690	0.3567	0.2269	0.2909	0.1686	0.3041	0.2208	0.2539
Gemini-3.1-Pro	0.1952	0.4275	0.2517	0.3321	0.1930	0.4084	0.2472	0.3179	0.1829	0.3280	0.2391	0.2754
GPT-5.4	**0.2207**	**0.4877**	**0.3093**	**0.4099**	0.1848	0.3948	**0.2670**	**0.3450**	0.1939	0.3539	**0.2717**	**0.3131**
Qwen3.5-Omni-Plus	0.2153	0.4746	0.2805	0.3710	**0.2017**	**0.4321**	0.2568	0.3341	**0.1961**	**0.3607**	0.2535	0.2950
Qwen3.5-Plus	0.2085	0.4589	0.2724	0.3603	0.1916	0.4088	0.2510	0.3252	0.1784	0.3268	0.2390	0.2779
Open-source Models												
Kimi-k2.5	0.1937	0.4271	0.2570	0.3415	0.1732	0.3694	0.2357	0.3055	0.1854	0.3393	0.2452	0.2837
Qwen3-VL-30B	0.1618	0.3579	0.2162	0.2875	0.1337	0.2888	0.1803	0.2345	0.1463	0.2670	0.1937	0.2241
Qwen3-VL-32B	0.1709	0.3783	0.2360	0.3135	0.1361	0.2918	0.1909	0.2471	0.1458	0.2678	0.1989	0.2302
Qwen3-VL-235B	0.1834	0.4067	0.2415	0.3216	0.1597	0.3465	0.2123	0.2783	0.1748	0.3208	0.2301	0.2680
Qwen3-Omni-Flash	0.1792	0.3965	0.2431	0.3232	0.1424	0.3066	0.2004	0.2615	0.1559	0.2882	0.2133	0.2497
Qwen3.5-27B	0.2043	0.4512	0.2598	0.3448	0.1876	0.3995	0.2430	0.3164	0.1841	0.3367	0.2380	0.2767
Qwen3.5-35B	0.1887	0.4160	0.2476	0.3277	0.1675	0.3584	0.2227	0.2878	0.1708	0.3110	0.2287	0.2649

aspects of model–human correspondence: agreement with population-level ratings and coverage of impression content reported by viewers.

4.3 Population-level Alignment across Human Rating and Disagreement

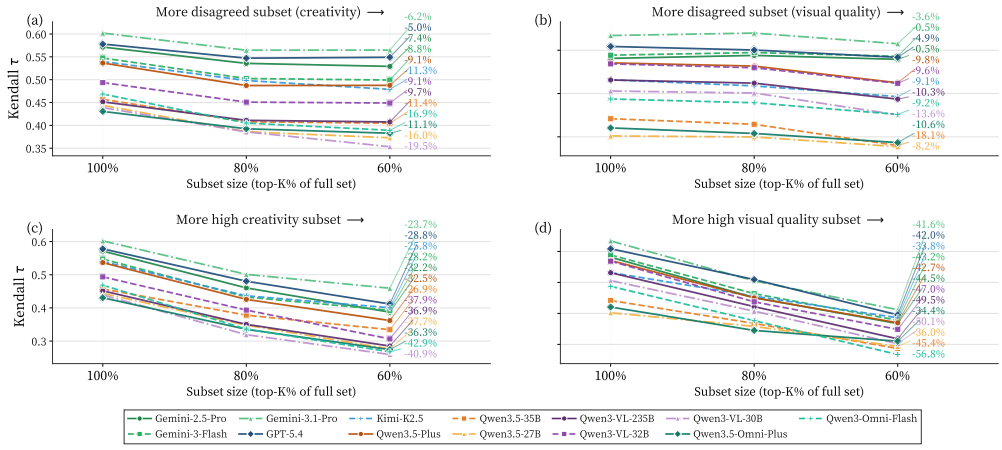


Figure 2: Alignment across progressively higher-rated and higher-disagreement video subsets.

We examine how population-level MLLM–human alignment varies across video subsets defined by human rating levels and disagreement. For each evaluation dimension, we separately rank videos by the mean or standard deviation of human ratings and progressively remove those with the lowest values. Mean-based selection therefore increasingly concentrates the retained set on highly rated videos, whereas standard-deviation-based selection concentrates it on videos with greater human disagreement. Alignment is recomputed within each retained subset using the original video-level MOS. Figure 2 presents Kendall's τ for representative creativity and visual perceptual quality dimensions. Pearson correlation and linear-weighted κ exhibit broadly similar trends across most models and dimensions; full results are provided in Appendix F.

Alignment generally declines under both selection schemes, with a more pronounced decrease in higher-rated subsets than in higher-disagreement subsets. In particular, models show weaker ranking correspondence with human judgments when distinguishing among videos rated highly for creativity or visual perceptual quality. This indicates that alignment observed across the full video set does not necessarily carry over to the high-rating end. Together, these findings highlight a boundary of current models' subjective rating alignment: their ability to track human judgments across a broad rating range does not consistently extend to fine-grained differentiation among highly rated videos, and alignment also weakens as viewer disagreement increases.

4.4 TRAIT-BASED VARIATION IN MLLM–HUMAN ALIGNMENT

Beyond population-level alignment, an important question is how model–human rating correspondence varies with viewer traits. Even when a model appears well aligned with humans on average, aggregate results may mask differences in its correspondence with viewers with different psychological traits. Using independently collected model and human ratings, we therefore test whether individual traits moderate MLLM–human rating correspondence under a generic viewer prompt without individual trait information. We focus on traits theoretically related to subjective responses during video viewing, including the Big Five (Rammstedt & John, 2007), regulatory focus (Higgins et al., 2001), consumer decision styles (Sprotles & Kendall, 1986; Sproles & Sproles, 1990), and emotional contagion (Doherty, 1997). Detailed subdimensions, questionnaire items, and sample distributions are reported in Appendix C. We fit the following mixed-effects moderation model:

We fit a mixed-effects moderation model, $S_{iv}^{(h)} = \beta_0 + \beta_1 S_v^{(a)} + \beta_2 C_i + \beta_3 (S_v^{(a)} C_i) + (1 \mid i) + (1 \mid v) + \epsilon_{iv}$, where $S_{iv}^{(h)}$ is participant i's rating of video v, averaged across relevant creativity or perceptual-quality subdimensions; $S_v^{(a)}$ is the corresponding MLLM rating averaged over three responses; and C_i is the participant's trait score. Both predictors are mean-centered. The model includes participant and video random intercepts. The interaction coefficient β_3 captures trait-related differences in model–human rating correspondence. Figure 3 presents results for the Big Five, regulatory focus, and emotional contagion; consumer decision-style results appear in Appendix G.

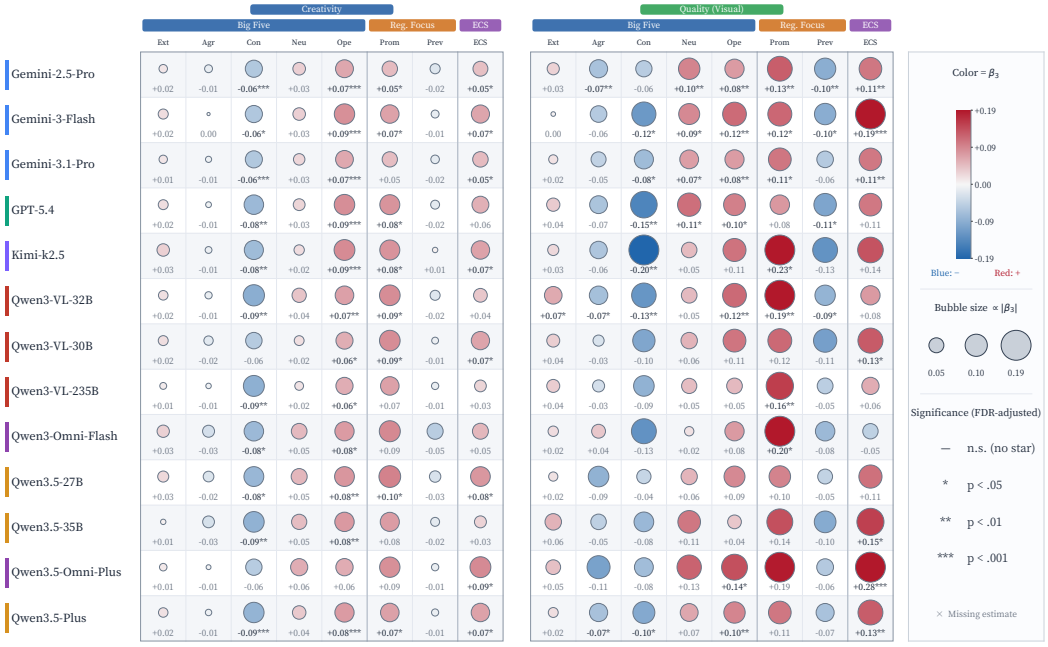


Figure 3: Mixed-effects interaction estimates (β_3) for Creativity and Visual Perceptual Quality. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (FDR-corrected).

Several recurring moderation patterns emerge across models. **Openness**, **promotion focus**, and **emotional contagion** generally exhibit positive interaction coefficients, whereas **conscientiousness** tends to exhibit negative coefficients. These patterns show that the same model ratings correspond differently to the judgments of viewers at different trait levels, even though the models receive no information about these traits. Consumer decision styles reveal additional variation, with generally positive moderation for **habit orientation** and **price/value orientation**, and negative moderation for **brand orientation**, **fashion orientation**, and **impulsive orientation**.

Taken together, these findings suggest that prompting a model to act as a generic viewer does not automatically make it representative of an "average viewer." Even without trait-specific instructions, its default ratings exhibit trait-related patterns of correspondence with human judgments that population-level alignment alone can obscure.

4.5 A Case Study of Information-processing Prompts and Alignment

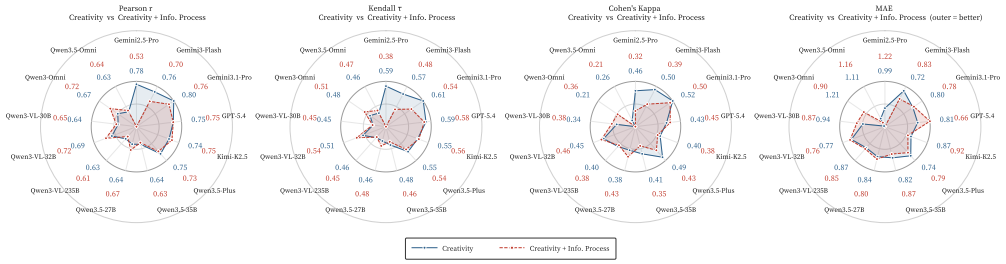


Figure 4: Performance change on creativity before and after adding information-processing.

We examine three illustrative cases selected based on their creativity-alignment profiles before and after adding information-processing questions: **Gemini-3.1-Pro** as a strong-performing reference, **Qwen3-Omni-Flash** as a case of improved alignment, and **Gemini-2.5-Pro** as a case of reduced alignment. The added questions cover **narrative**, **desire to postpone closure** (DPC), **positive affect**, and **attention capture**. Figure 4 compares alignment on the same creativity items across prompt settings; Table 4 reports correspondence on the added items; and Fig. 26 compares human and model correlation and partial-correlation networks. Full results are provided in Appendix H.

On the added items, narrative generally shows stronger correspondence than affective responses such as interested, inspired, and pleasant, while several DPC items show lower correspondence. Gemini-3.1-Pro performs strongly across many items. Among the two cases with contrasting changes in creativity alignment, Qwen3-Omni-Flash shows higher correspondence than Gemini-2.5-Pro on most added items. These cases illustrate how changes in creativity alignment can be examined alongside models' correspondence with human information-processing reports.

Table 4: MLLM–human correspondence on Creativity and process dimensions for the case models.

Metric	Model	Creativity								Information Processing										
		Divergence					General			Narrative		Attn.Catch		DPC			Positive Affect			
		Orig.	Flex.	Synth.	Elab.	Art.	OvCr	VACoord	ProdQ	Sty	Plot	Hook	UnExp	DPC-1	DPC-2	DPC-3	Int.	Surp.	Insp.	Pleas.
Pear Corr.	Gemini-2.5-Pro	0.557	0.596	0.577	0.417	0.476	0.563	0.427	0.514	0.719	0.720	0.534	0.576	0.386	−0.172	0.171	0.390	0.600	0.347	0.322
	Gemini-3.1-Pro	**0.787**	**0.799**	**0.779**	**0.700**	**0.711**	**0.807**	**0.632**	**0.735**	0.803	0.795	**0.704**	**0.765**	**0.592**	0.152	**0.337**	**0.682**	**0.749**	**0.560**	**0.621**
	Qwen3-Omni-Flash	0.730	0.726	0.763	0.647	0.702	0.744	0.631	0.637	**0.807**	**0.826**	0.620	0.705	0.528	**0.350**	0.122	0.582	0.708	0.420	0.488
Kendall's τ	Gemini-2.5-Pro	0.424	0.418	0.436	0.274	0.304	0.429	0.277	0.339	0.586	0.600	0.331	0.436	0.321	−0.084	0.128	0.288	0.450	0.214	0.231
	Gemini-3.1-Pro	**0.581**	**0.555**	**0.582**	0.446	0.472	**0.600**	**0.467**	**0.492**	**0.635**	0.630	**0.481**	**0.588**	**0.422**	0.098	**0.248**	**0.511**	**0.590**	**0.404**	**0.449**
	Qwen3-Omni-Flash	0.514	0.498	0.556	**0.480**	**0.507**	0.527	0.436	0.455	0.634	**0.653**	0.427	0.530	0.365	**0.247**	0.079	0.426	0.548	0.238	0.366
Cohen's Kappa	Gemini-2.5-Pro	0.347	0.363	0.321	0.269	0.266	0.362	0.021	0.134	0.381	0.350	0.169	0.338	0.168	−0.056	0.030	0.149	0.331	0.106	0.066
	Gemini-3.1-Pro	**0.520**	**0.533**	**0.514**	**0.451**	**0.455**	**0.547**	**0.205**	**0.348**	0.501	0.479	**0.460**	**0.467**	0.249	0.007	**0.187**	**0.391**	**0.424**	**0.221**	**0.230**
	Qwen3-Omni-Flash	0.429	0.325	0.513	0.230	0.332	0.346	0.150	0.100	**0.545**	**0.568**	0.165	0.466	**0.292**	**0.056**	0.051	0.075	0.376	0.170	0.041

The networks provide a descriptive view of these differences. In the human partial-correlation network, creativity is closely linked to general evaluation and attention capture, but less strongly linked to DPC. Gemini-3.1-Pro shows several relationships similar to those in the human network while keeping DPC relatively distinct. Qwen3-Omni-Flash differs from the human pattern in some relationships, whereas Gemini-2.5-Pro shows more pronounced departures. Together, the three cases show different creativity-alignment changes across prompt settings, accompanied by differences in correspondence with human information-processing reports and in response-level correlation patterns. Considering these results together provides a fuller picture of each model's response profile than creativity alignment alone.

5 Conclusion

We introduce **AdAlign**, a benchmark for subjective video understanding that combines population-level rating evaluation, open-ended impressions, and individual-level analyses. Current MLLMs capture broad human rating trends but exhibit rating discrepancies and weaker ranking correspondence in higher-rated and higher-disagreement subsets. Trait analyses reveal distinct correspondence patterns under generic viewer prompts. An exploratory case study suggests that information-processing prompts may support creativity alignment when responses correspond well with human reports, though benefits vary across models. These findings motivate evaluation and modeling attentive to both population-level judgments and the diversity of viewers' subjective experiences.

AI USE STATEMENT

Generative AI tools were used for language editing, improving the clarity and organization of the manuscript, brainstorming alternative ways to present and discuss the authors' existing analyses, and assisting with the creation and editing of software code. All AI-assisted code was reviewed and tested for correctness by two authors. Generative AI tools were not used for data collection or generation, annotation, research methodology or experimental design, experiment execution, quantitative analysis, or result interpretation. All scientific decisions, analyses, and conclusions were made and verified by the authors. The authors critically reviewed and revised all AI-assisted content and take full responsibility for the final manuscript and accompanying artifacts.

ETHICS STATEMENT

This study involved human participants and was reviewed and approved by the appropriate ethics review body at the authors' institution before data collection. All participants provided informed consent. Participant data were de-identified before analysis, and only de-identified participant information is included in the released dataset. We took precautions throughout data processing, reporting, and release to protect participant privacy and reduce the risk of re-identification.

AdAlign is intended to support the evaluation of MLLM–human correspondence in subjective video understanding, rather than to define a universal or normative standard of viewer judgment. Its participant sample and advertising context may not represent viewers from other demographic or cultural settings. Moreover, correspondence with aggregate human ratings does not mean that model outputs represent every viewer or can substitute for individual human judgments. Trait-related results characterize aggregate patterns of correspondence under the evaluated generic-viewer prompts and should not be used to infer or profile the psychological traits of individual viewers. In applications such as advertising assessment, content evaluation, and creative support, model judgments should therefore complement rather than replace evaluations from relevant human audiences.

REPRODUCIBILITY STATEMENT

We provide detailed experimental settings, model configurations, prompt designs, and evaluation procedures in the main paper and appendices. The anonymized implementation code is available at `https://anonymous.4open.science/r/AdAlign_Bench`, and the de-identified benchmark data are available in supplementary material. These resources are provided to support reproduction of the benchmark evaluation and analyses reported in this paper.

REFERENCES

Jennifer L Aaker and Angela Y Lee. "i" seek pleasures and "we" avoid pains: The role of self-regulatory goals in information processing and persuasion. *Journal of consumer Research*, 28(1): 33–49, 2001.

Arctanx An, Shizhao Sun, Danqing Huang, Mingxi Cheng, Yan Gao, Ji Li, Yu Qiao, and Jiang Bian. Can vision language models assess graphic design aesthetics? a benchmark, evaluation, and dataset perspective. *arXiv preprint arXiv:2603.01083*, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 961–970, 2015.

Giovanni Luca Cascio Rizzo, Jonah Berger, Matteo De Angelis, and Rumen Pozharliev. How sensory language shapes influencer's impact. *Journal of Consumer Research*, 50(4):810–825, 2023.

Xing-Yu Chu, Chun-Tuan Chang, and Angela Y Lee. Values created from far and near: Influence of spatial distance on brand evaluation. *Journal of Marketing*, 85(6):162–175, 2021.

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

Google Deepmind. Gemini 3 flash: frontier intelligence built for speed, December 2025. URL https://blog.google/products-and-platforms/products/gemini/gemini-3-flash/.

Google Deepmind. Gemini 3.1 pro: A smarter model for your most complex tasks, February 2026. URL https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/.

R William Doherty. The emotional contagion scale: A measure of individual differences. *Journal of nonverbal Behavior*, 21(2):131–154, 1997.

Ryan S Elder and Aradhna Krishna. The "visual depiction effect" in advertising: Facilitating embodied mental simulation through product orientation. *Journal of Consumer Research*, 38(6): 988–1003, 2012.

Jennifer Edson Escalas. Narrative processing: Building consumer connections to brands. *Journal of consumer psychology*, 14(1-2):168–180, 2004.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24108–24118, 2025.

Melanie C Green and Timothy C Brock. The role of transportation in the persuasiveness of public narratives. *Journal of personality and social psychology*, 79(5):701, 2000.

Linda Hagen. Pretty healthy food: How and when aesthetics enhance perceived healthiness. *Journal of Marketing*, 85(2):129–145, 2021.

E Tory Higgins, Ronald S Friedman, Robert E Harlow, Lorraine Chen Idson, Ozlem N Ayduk, and Amy Taylor. Achievement orientations from subjective histories of success: Promotion pride versus prevention pride. *European Journal of Social Psychology*, 31(1):3–23, 2001.

Jacob B Hirsh, Sonia K Kang, and Galen V Bodenhausen. Personalized persuasion: Tailoring persuasive appeals to recipients' personality traits. *Psychological science*, 23(6):578–581, 2012.

Xun Huang and Aparna A Labroo. Cueing morality: The effect of high-pitched music on healthy choice. *Journal of Marketing*, 84(6):130–143, 2020.

He Jia, B Kyu Kim, and Lin Ge. Speed up, size down: How animated movement speed in product videos influences size assessment and product evaluation. *Journal of Marketing*, 84(5):100–116, 2020.

Qi Jia, Baoyu Fan, Cong Xu, Lu Liu, Liang Jin, Guoguang Du, Zhenhua Guo, Yaqian Zhao, Xuanjing Huang, and Rengang Li. Infer induced sentiment of comment response to video: A new task, dataset and baseline. *Advances in Neural Information Processing Systems*, 37:103737–103750, 2024.

Ruixiang Jiang and Chang Wen Chen. Multimodal llms can reason about aesthetics in zero-shot. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 6634–6643, 2025.

Frank R Kardes, Maria L Cronley, James J Kellaris, and Steven S Posavac. The role of selective information processing in price-quality inference. *Journal of consumer research*, 31(2):368–374, 2004.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pp. 706–715, 2017.

Dasong Li, Wenjie Li, Baili Lu, Hongsheng Li, Sizhuo Ma, Gurunandan Krishnan, and Jian Wang. Delving deep into engagement prediction of short videos. In *European Conference on Computer Vision*, pp. 289–306. Springer, 2024a.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024b.

Yiyi Li and Ying Xie. Is a picture worth a thousand words? an empirical study of image content and social media engagement. *Journal of marketing research*, 57(1):1–19, 2020.

Zeqian Li, Shangzhe Di, Zhonghua Zhai, Weilin Huang, Yanfeng Wang, and Weidi Xie. Universal video temporal grounding with generative multi-modal large language models. *arXiv preprint arXiv:2506.18883*, 2025.

Baoyu Liang, Qile Su, Shoutai Zhu, Yuchen Liang, and Chao Tong. Videvent: a large dataset for understanding dynamic evolution of events in videos. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, pp. 5128–5136, 2025.

Zhaoyu Liu, Kan Jiang, Murong Ma, Zhe Hou, Yun Lin, and Jin Song Dong. F^3 set: Towards analyzing fast, frequent, and fine-grained events from videos. *arXiv preprint arXiv:2504.08222*, 2025.

George Loewenstein. The psychology of curiosity: A review and reinterpretation. *Psychological bulletin*, 116(1):75, 1994.

Xinwei Long, Kai Tian, Peng Xu, Guoli Jia, Jingxuan Li, Sa Yang, Yihua Shao, Kaiyan Zhang, Che Jiang, Hao Xu, et al. Adsqa: Towards advertisement video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23396–23407, 2025.

Deborah J MacInnis and C Whan Park. The differential role of characteristics of music on high-and low-involvement consumers' processing of ads. *Journal of consumer Research*, 18(2):161–173, 1991.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

Sandra C Matz, Michal Kosinski, Gideon Nave, and David J Stillwell. Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the national academy of sciences*, 114(48):12714–12719, 2017.

Grant McCracken. Culture and consumption: A theoretical account of the structure and movement of the cultural meaning of consumer goods. *Journal of consumer research*, 13(1):71–84, 1986.

Joan Meyers-Levy and Laura A Peracchio. Getting an angle in advertising: The effect of camera angle on product evaluations. *Journal of marketing research*, 29(4):454–461, 1992.

Lauren Min, Peggy J Liu, and Cary L Anderson. The visual complexity= higher production cost lay belief. *Journal of Consumer Research*, 51(6):1167–1185, 2025.

Asmar Nadeem, Faegheh Sardari, Robert Dawes, Syed Sameed Husain, Adrian Hilton, and Armin Mustafa. Narrativebridge: Enhancing video captioning with causal-temporal narrative. *arXiv preprint arXiv:2406.06499*, 2024.

Ellie Pavlick and Tom Kwiatkowski. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694, 2019.

Beatrice Rammstedt and Oliver P John. Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of research in Personality*, 41(1):203–212, 2007.

Anne L Roggeveen, Dhruv Grewal, Claudia Townsend, and Ram Krishnan. The impact of dynamic presentation format on consumer preferences for hedonic products and services. *Journal of Marketing*, 79(6):34–49, 2015.

Linda M Scott. Images in advertising: The need for a theory of visual rhetoric. *Journal of consumer research*, 21(2):252–273, 1994.

Julio Sevilla and Robert J Meyer. Leaving something for the imagination: The effect of visual concealment on preferences. *Journal of Marketing*, 84(4):109–126, 2020.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Robert E Smith, Scott B MacKenzie, Xiaojing Yang, Laura M Buchholz, and William K Darley. Modeling the determinants and effects of creativity in advertising. *Marketing science*, 26(6): 819–833, 2007.

Elizabeth Kendall Sproles and George B Sproles. Consumer decision-making styles as a function of individual learning styles. *Journal of Consumer Affairs*, 24(1):134–147, 1990.

George B Sprotles and Elizabeth L Kendall. A methodology for profiling consumers' decision-making styles. *Journal of Consumer Affairs*, 20(2):267–279, 1986.

Shangkun Sun, Xiaoyu Liang, Songlin Fan, Wenxu Gao, and Wei Gao. Ve-bench: subjective-aligned benchmark suite for text-driven video editing quality assessment. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, pp. 7105–7113, 2025.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, et al. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

Qwen Team. Qwen3.5-omni: Scaling up, toward native omni-modal agi, March 2026a. URL `https://qwen.ai/blog?id=qwen3.5-omni`.

Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents, February 2026b. URL https://qwen.ai/blog?id=qwen3.5.

Thales Teixeira, Rosalind Picard, and Rana El Kaliouby. Why, when, and how much to entertain consumers in advertisements? a web-based facial tracking field study. *Marketing Science*, 33(6): 809–827, 2014.

Gerard J Tellis, Deborah J MacInnis, Seshadri Tirunillai, and Yanwei Zhang. What drives virality (sharing) of online digital content? the critical role of information, emotion, and brand prominence. *Journal of marketing*, 83(4):1–20, 2019.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72: 1385–1470, 2021.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4581–4591, 2019.

Qi Wu, Quanlong Zheng, Yanhao Zhang, Junlin Xie, Jinguo Luo, Kuo Wang, Peng Liu, Qingsong Xie, Ru Zhen, Zhenyu Yang, et al. H2vu-benchmark: A comprehensive benchmark for hierarchical holistic video understanding. *arXiv preprint arXiv:2503.24008*, 2025.

Yupeng Xie, Zhiyang Zhang, Yifan Wu, Sirong Lu, Jiayi Zhang, Zhaoyang Yu, Jinlin Wang, Sirui Hong, Bang Liu, Chenglin Wu, et al. Visjudge-bench: Aesthetics and quality assessment of visualizations. *arXiv preprint arXiv:2510.22373*, 2025.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025a.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5288–5296, 2016.

Yijie Xu, Bolun Zheng, Wei Zhu, Hangjia Pan, Yuchen Yao, Ning Xu, Anan Liu, Quan Zhang, and Chenggang Yan. Smtpd: A new benchmark for temporal prediction of social media popularity. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18847–18857, 2025b.

Xiaojing Yang and Robert E Smith. Beyond attention effects: Modeling the persuasive and emotional effects of advertising creativity. *Marketing Science*, 28(5):935–949, 2009.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: a dataset for understanding complex web videos via question answering. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, pp. 9127–9134, 2019.

Chenkai Zhang, Yiming Lei, Zeming Liu, Haitao Leng, ShaoGuo Liu, Tingting Gao, Qingjie Liu, and Yunhong Wang. Seriesbench: A benchmark for narrative-driven drama series understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28995–29004, 2025.

APPENDIX

A LIMITATIONS

This work has three main limitations. First, the participant sample is concentrated primarily in the 20–30 age range because of the constraints of in-person data collection. The findings may therefore not generalize to viewers from other age groups, occupational backgrounds, cultural contexts, or patterns of social-media use. Random video assignment supports the internal validity of comparisons within the collected sample, but does not resolve this limitation in demographic representativeness. Future work could extend the benchmark to more diverse participant populations and examine whether the observed alignment patterns generalize across demographic and cultural settings.

Second, AdAlign is designed primarily as an evaluation and analysis benchmark rather than a large-scale training resource. Although its 225 participants and 393 videos support the population-level, trait-based, and process-level analyses conducted in this work, the number of responses available for each video limits fine-grained estimation of the full distribution of viewer judgments. The current scale is also insufficient for training-intensive settings such as supervised fine-tuning, preference learning, or reinforcement learning. Larger-scale data collection would support more detailed analyses of individual variation and the development of training methods for subjective video understanding.

Third, the current evaluation does not fully disentangle how different input modalities contribute to model judgments. AdAlign includes both visual and audio-related items, but performance on these items alone cannot determine whether a model directly uses audio information or infers related judgments from visual and linguistic cues. Our results therefore characterize alignment under the evaluated model and prompt settings rather than establishing how individual modalities produce that alignment. More controlled evaluations are needed to isolate the contributions of visual, audio, and cross-modal information to subjective video judgments.

B PARTICIPANT ANNOTATION INTERFACE

This section describes the participant annotation procedure and illustrates the main pages of the human annotation interface, including the welcome page, video viewing page, and questionnaire page (Figures 5–7). Before the study began, all participants provided informed consent, confirming that their data would be used only for research purposes and that no private information would be disclosed. After entering the study, participants first read the task instructions and then completed practice tasks in the same format as the formal ones, in order to familiarize themselves with the procedure and response format.

For each video, participants were required to watch it in full at least once before answering the questions, and the corresponding product name was revealed only after one complete viewing to prevent prior product information from influencing the viewing experience. Once participants entered the questionnaire page for a given video, replay was not allowed. They could proceed to the next video only after finishing all questions for the current one. Each participant completed approximately 20–25 videos. After all video tasks were finished, participants completed a post-task questionnaire that included personality trait scales. Participants were recruited from student and working-adult populations, and none had received professional annotation training.

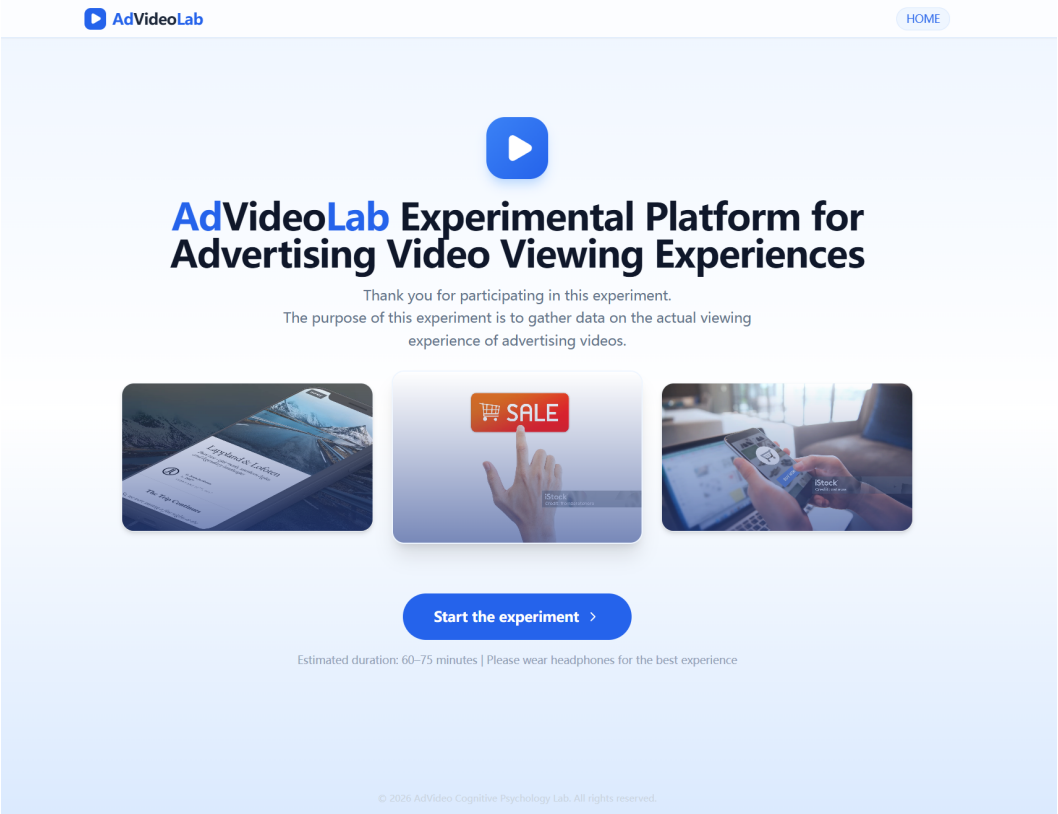


Figure 5: Welcome Page of the Human Annotation Interface

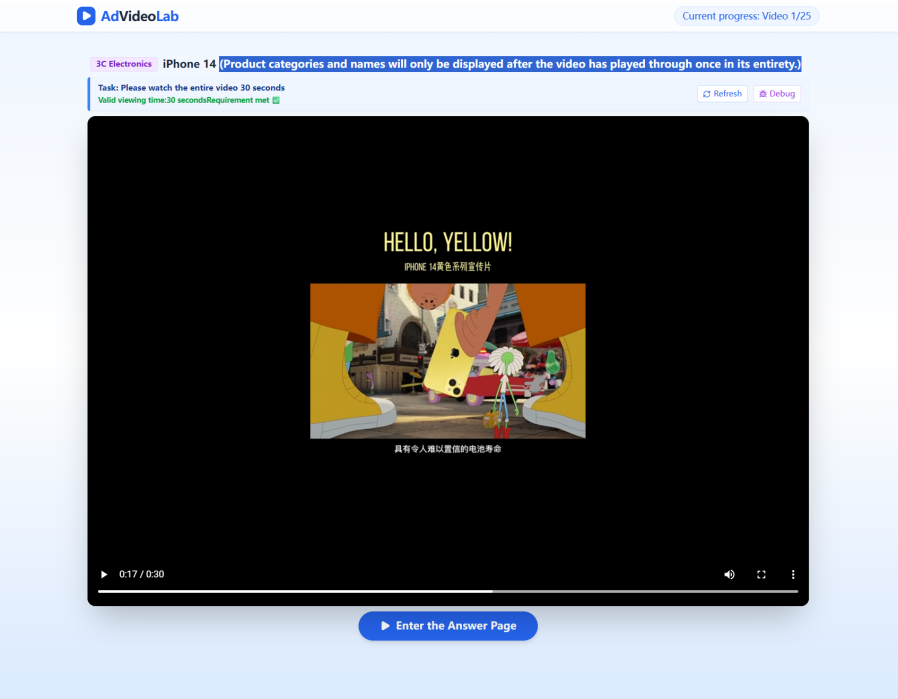


Figure 6: Video Viewing Page of the Human Annotation Interface

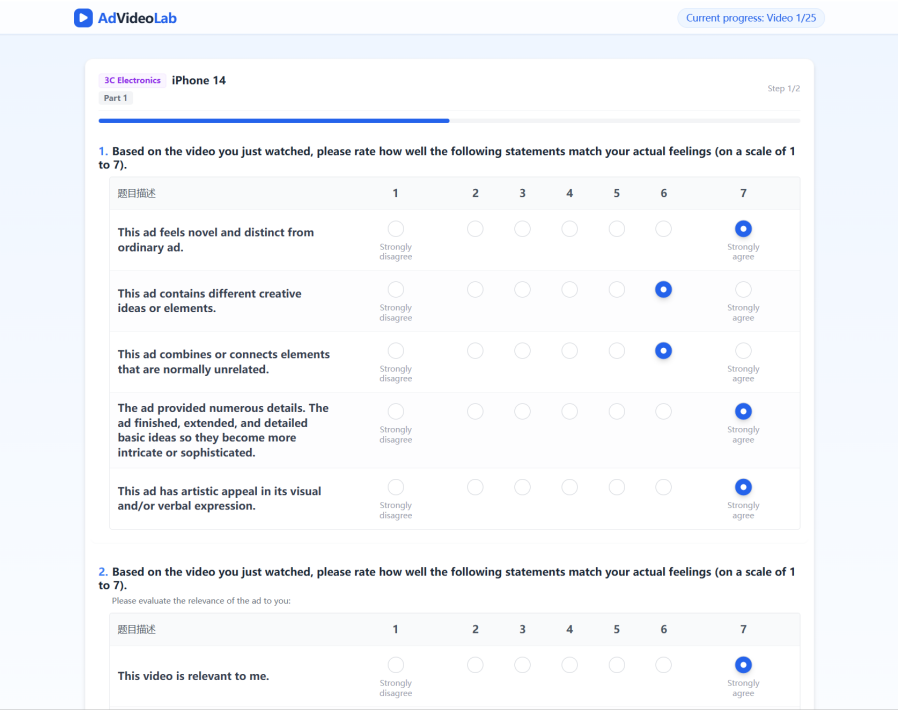


Figure 7: Questionnaire Page of the Human Annotation Interface

B.1 ANNOTATION ITEMS

This section presents the exact items shown to participants during annotation. To preserve the naturalness of subjective judgment as much as possible, participants saw only the item descriptions during the annotation process, but not the corresponding higher-level dimensions, subdimensions, or their formal definitions. For example, participants did not see labels such as creativity or multimodal perceptual quality, nor did they see whether a given item was intended to assess a specific subdimension such as originality or artistic value. They were also not provided with textual definitions of these constructs. This design was intended to reduce potential framing effects introduced by predefined category labels or measurement constructs, and to encourage judgments grounded in participants' actual viewing experiences.

Regarding the response scales, positive affect was rated on a 1–5 Likert scale, while all other items used a 1–7 Likert scale. We provide the full list of item descriptions in the Table 5 to support transparency and reproducibility of the annotation process.

Table 5: Itmes for Human & MLLMs Evaluation.

Item	Item Description
Content Evaluation	
Creativity (Smith et al., 2007; Yang & Smith, 2009)	
Originality	This ad feels novel and distinct from ordinary ad.
Flexibility	This ad contains different creative ideas or elements.
Synthesis	This ad combines or connects elements that are normally unrelated.
Elaboration	The ad provided numerous details. The ad finished, extended, and detailed basic ideas so they become more intricate or sophisticated.
Artistic Value	This ad has artistic appeal in its visual and/or verbal expression.
Overall Creativity	Overall, this ad is very creative.
Multimodal Perceptual Quality(Visual)	
Complexity	The video contains rich visual elements in its composition.
Technique	The video demonstrates highly professional shooting techniques.
Flow	The video has a strong sense of visual flow in its editing and pacing.
Aesthetic Appeal	The visuals of this video have high aesthetic appeal.
Visual Pleasantness	The visuals of this video have high visual pleasantness.
Visual Attractiveness	The visuals of this video have high visual attractiveness.
Multimodal Perceptual Quality(Audio)	
Has BGM?	Does this video contain background music (BGM)?
BGM–Visual Match	The background music matches the video content well.
Has Speech?	Does this video contain narration or spoken voice? (excluding sung vocals in the background music)
Speech Comfort	The narration or spoken voice is comfortable to listen to.
Speech Fit	The narration or spoken voice matches the product and/or ad well.
Speech Professionalism	The narration or spoken explanation sounds professional.
Multimodal Perceptual Quality(Overall)	
VA Coordination	Overall, the visuals and sound of this video are well coordinated.
Production Quality	Overall, this video has high production quality.
Information Processing	
Attention Catch	
Hook	The opening of this video is very engaging and makes me want to keep watching.
Unexpected	The content of this video felt unexpected to me.
Narrative	
Story	I think this video has a strong sense of story.
Plot	I think the content of this video has a clear plot.
DPC (Desire to Postpone Closure)	
DPC-1	While watching, I did not want to form a judgment about the brand/product too quickly, and instead wanted to see what additional information would follow.
DPC-2	I considered both the potential benefits and drawbacks of the brand/product rather than rushing to a judgment.
DPC-3	I felt that this short video did not provide enough information for me to make a reliable judgment.
Positive Affect	
Interested	Watching this video makes you feel Interested.
Surprised	Watching this video makes you feel Surprised.
Inspired	Watching this video makes you feel Inspired.
Pleasant	Watching this video makes you feel Pleasant.
Open Impression	
Open-end Item	What specific content in this video left a strong impression on you? (You can list multiple). For the content you mentioned, is your impression positive or negative?

C PERSONALITY-RELATED TRAIT MEASUREMENT

After completing the video annotation tasks, participants were asked to fill out a post-task questionnaire. The questionnaire contained basic information(age range and gender) established scales from prior literature and did not collect sensitive personal information. We collected several individual-difference measures, including the Big Five personality traits, regulatory focus, consumer decision-making styles, and emotional contagion. The full list of scale items is provided below.

During questionnaire administration, participants saw only the text of each item. They did not see reverse-scoring markers (e.g., (R)) or the dimension to which each item belonged. This design was intended to minimize the influence of scale-structure information on responses and to better capture participants' natural self-reports.

C.1 BIG FIVE PERSONALITY TRAITS

The Big Five is a widely used framework in personality psychology, consisting of five broad dimensions: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness. In this work, we used the Big Five Inventory-10 (BFI-10) as a brief measure of participants' Big Five traits. Participants responded to the instruction, *"How well do the following statements describe your personality?"*, on a 5-point scale ranging from (1) *Disagree strongly* to (5) *Agree strongly*. The BFI-10 dimensions were scored as follows: Extraversion (Items 1R, 6), Agreeableness (Items 2, 7R), Conscientiousness (Items 3R, 8), Neuroticism (Items 4R, 9), and Openness (Items 5R, 10), where R indicates reverse-scored items. Detailed questionnaire items are shown in Table 6.

C.2 REGULATORY FOCUS

Regulatory focus describes two fundamental self-regulatory orientations in goal pursuit: promotion focus and prevention focus. Promotion focus emphasizes growth, accomplishment, and the pursuit of ideals, whereas prevention focus emphasizes duties, responsibilities, and the avoidance of mistakes or losses. In this work, we used the Regulatory Focus Questionnaire (RFQ) to measure participants' regulatory focus. The scale consists of two dimensions, promotion focus and prevention focus. All items were rated on a 5-point scale, and items marked with (R) are reverse-scored. Detailed questionnaire items are shown in Table 7.

C.3 CONSUMER DECISION-MAKING STYLES

Consumer decision-making styles describe relatively stable orientations that consumers adopt when making purchase decisions. In this work, we used the Consumer Style Inventory (CSI) to measure participants' consumer decision-making styles. The scale consists of eight dimensions: perfectionistic/high-quality consciousness, brand consciousness, novelty-fashion consciousness, recreational/hedonistic shopping orientation, price/value consciousness, impulsive/careless shopping orientation, confusion from overchoice, and habitual/brand-loyal orientation. All items were rated on a 5-point scale, and items marked with (R) are reverse-scored. In the main analysis, these dimensions are examined separately rather than aggregated into a single overall score, as they capture distinct consumer-style tendencies that may relate differently to subjective video judgments. Detailed questionnaire items are shown in Table 8.

Table 6: Questionnaire Items for the Big Five Inventory-10 (BFI-10)

I see myself as someone who...	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
1. ...is reserved (R)	◯	◯	◯	◯	◯
2. ...is generally trusting	◯	◯	◯	◯	◯
3. ...tends to be lazy (R)	◯	◯	◯	◯	◯
4. ...is relaxed, handles stress well (R)	◯	◯	◯	◯	◯
5. ...has few artistic interests (R)	◯	◯	◯	◯	◯
6. ...is outgoing, sociable	◯	◯	◯	◯	◯
7. ...tends to find fault with others (R)	◯	◯	◯	◯	◯
8. ...does a thorough job	◯	◯	◯	◯	◯
9. ...gets nervous easily	◯	◯	◯	◯	◯
10. ...has an active imagination	◯	◯	◯	◯	◯

Table 7: Questionnaire Items for the Regulatory Focus Questionnaire (RFQ)

Item	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
Promotion Focus					
Compared to most people, are you typically unable to get what you want out of life? (R)	◯	◯	◯	◯	◯
How often have you accomplished things that got you "psyched" to work even harder?	◯	◯	◯	◯	◯
Do you often do well at different things that you try?	◯	◯	◯	◯	◯
When it comes to achieving things that are important to me, I find that I don't perform as well as I ideally would like to do. (R)	◯	◯	◯	◯	◯
I feel like I have made progress toward being successful in my life.	◯	◯	◯	◯	◯
I have found very few hobbies or activities in my life that capture my interest or motivate me to put effort into them. (R)	◯	◯	◯	◯	◯
Prevention Focus					
Growing up, would you ever "cross the line" by doing things that your parents would not tolerate? (R)	◯	◯	◯	◯	◯
Did you get on your parents' nerves often when you were growing up? (R)	◯	◯	◯	◯	◯
How often did you obey rules and regulations that were established by your parents?	◯	◯	◯	◯	◯
Growing up, did you ever act in ways that your parents thought were objectionable? (R)	◯	◯	◯	◯	◯
Not being careful enough has gotten me into trouble at times. (R)	◯	◯	◯	◯	◯

Table 8: Questionnaire Items for the Consumer Style Inventory (CSI)

Factors and Items	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
Factor 1: Perfectionistic, High-Quality Conscious Consumer					
Getting very good quality is very important to me.	○	○	○	○	○
When it comes to purchasing products, I try to get the very best or perfect choice.	○	○	○	○	○
In general, I usually try to buy the best overall quality.	○	○	○	○	○
I make special effort to choose the very best quality products.	○	○	○	○	○
I really don't give my purchases much thought or care. (R)	○	○	○	○	○
My standards and expectations for products I buy are very high.	○	○	○	○	○
I shop quickly, buying the first product or brand I find that seems good enough. (R)	○	○	○	○	○
A product doesn't have to be perfect, or the best, to satisfy me. (R)	○	○	○	○	○
Factor 2: Brand Conscious, "Price Equals Quality" Consumer					
The well-known national brands are best for me.	○	○	○	○	○
The more expensive brands are usually my choices.	○	○	○	○	○
The higher the price of a product, the better its quality.	○	○	○	○	○
Nice department and specialty stores offer me the best products.	○	○	○	○	○
I prefer buying the best-selling brands.	○	○	○	○	○
The most advertised brands are usually very good choices.	○	○	○	○	○
Factor 3: Novelty-Fashion Conscious Consumer					
I usually have one or more outfits of the very newest style.	○	○	○	○	○
I keep my wardrobe up-to-date with the changing fashions.	○	○	○	○	○
Fashionable, attractive styling is very important to me.	○	○	○	○	○
To get variety, I shop different stores and choose different brands.	○	○	○	○	○
It's fun to buy something new and exciting.	○	○	○	○	○
Factor 4: Recreational, Hedonistic Consumer					
Shopping is not a pleasant activity to me. (R)	○	○	○	○	○
Going shopping is one of the enjoyable activities of my life.	○	○	○	○	○
Shopping the stores wastes my time. (R)	○	○	○	○	○
I enjoy shopping just for the fun of it.	○	○	○	○	○
I make my shopping trips fast. (R)	○	○	○	○	○
Factor 5: Price Conscious, "Value for Money" Consumer					
I buy as much as possible at sale prices.	○	○	○	○	○
The lower price products are usually my choice.	○	○	○	○	○
I look carefully to find the best value for the money.	○	○	○	○	○
Factor 6: Impulsive, Careless Consumer					
I should plan my shopping more carefully than I do.	○	○	○	○	○
I am impulsive when purchasing.	○	○	○	○	○
Often I make careless purchases I later wish I had not.	○	○	○	○	○
I take the time to shop carefully for best buys. (R)	○	○	○	○	○
I carefully watch how much I spend. (R)	○	○	○	○	○
Factor 7: Confused by Overchoice Consumer					
There are so many brands to choose from that often I feel confused.	○	○	○	○	○
Sometimes it's hard to choose which stores to shop.	○	○	○	○	○
The more I learn about products, the harder it seems to choose the best.	○	○	○	○	○
All the information I get on different products confuses me.	○	○	○	○	○
Factor 8: Habitual, Brand-Loyal Consumer					
I have favorite brands I buy over and over.	○	○	○	○	○
Once I find a product or brand I like, I stick with it.	○	○	○	○	○
I go to the same stores each time I shop.	○	○	○	○	○
I change brands I buy regularly. (R)	○	○	○	○	○

C.4 EMOTIONAL CONTAGION (EC) SCALE

Emotional contagion refers to the tendency to automatically perceive, resonate with, and internalize the emotional states of others during social interaction. In this work, we used the Emotional Contagion (EC) Scale to measure participants' susceptibility to emotional contagion. The scale covers five emotion-related subdimensions: happiness, love, fear, anger, and sadness. All items were rated on a 5-point scale. In the main analysis, we do not treat these subdimensions separately; instead, we average responses across all items to obtain an overall measure of emotional contagion. Detailed questionnaire items are shown in Table 9.

All items were rated on a 5-point Likert scale: (1) Never, (2) Rarely, (3) Usually, (4) Often, (5) Always.

In our analysis, we do not treat these subdimensions separately. Instead, we aggregate responses across all subdimensions and compute their average as an overall measure of participants' emotional contagion. As with the other personality-related trait questionnaires, participants saw only the item text during administration and were not informed of the subdimension associated with each item.

Table 9: Emotional Contagion (EC) Scale

Item	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
Happiness					
Being with a happy person picks me up when I'm feeling down.	○	○	○	○	○
When someone smiles warmly at me, I smile back and feel warm inside.	○	○	○	○	○
Being around happy people fills my mind with happy thoughts.	○	○	○	○	○
Love					
When I look into the eyes of the one I love, my mind is filled with thoughts of romance.	○	○	○	○	○
I melt when the one I love holds me close.	○	○	○	○	○
I sense my body responding when the one I love touches me.	○	○	○	○	○
Fear					
Watching the fearful faces of victims on the news makes me try to imagine how they might be feeling.	○	○	○	○	○
I notice myself getting tense when I'm around people who are stressed out.	○	○	○	○	○
Listening to the shrill screams of a terrified child in a dentist's waiting room makes me feel nervous.	○	○	○	○	○
Anger					
I clench my jaws and my shoulders get tight when I see the angry faces on the news.	○	○	○	○	○
It irritates me to be around angry people.	○	○	○	○	○
I tense when overhearing an angry quarrel.	○	○	○	○	○
Sadness					
If someone I'm talking with begins to cry, I get teary-eyed.	○	○	○	○	○
I get filled with sorrow when people talk about the death of their loved ones.	○	○	○	○	○
I cry at sad movies.	○	○	○	○	○

D SUPPLEMENTARY DATASET DESCRIPTION

D.1 DISTRIBUTION OF PARTICIPANTS

This section provides additional statistics on participant characteristics and rating distributions. In total, we recruited 225 participants, whose gender and age distributions are shown in Figure 8. The sample mainly consisted of students and working adults, with participants concentrated in the young-adult age range (approximately 20–30 years old). We acknowledge that, due to the constraints of in-person data collection, the current sample does not fully match the demographic composition of the broader population. This limitation also points to an important direction for future dataset expansion: improving the coverage and representativeness of participant sources.

At the same time, the current sample still captures a meaningful segment of real-world viewers and supports informative analysis of subjective judgment variation within this population. In particular, we made efforts to ensure that the collected sample exhibits a reasonable distribution with respect to the personality-related measures examined in this work, as shown in Figures 8 9. This allows us to study individual differences and subgroup-specific variation in MLLM–human alignment, while leaving broader demographic coverage to future work.

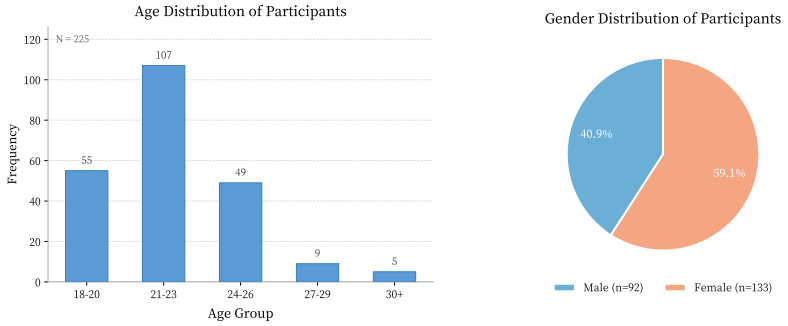


Figure 8: Distribution of participants age range and gender

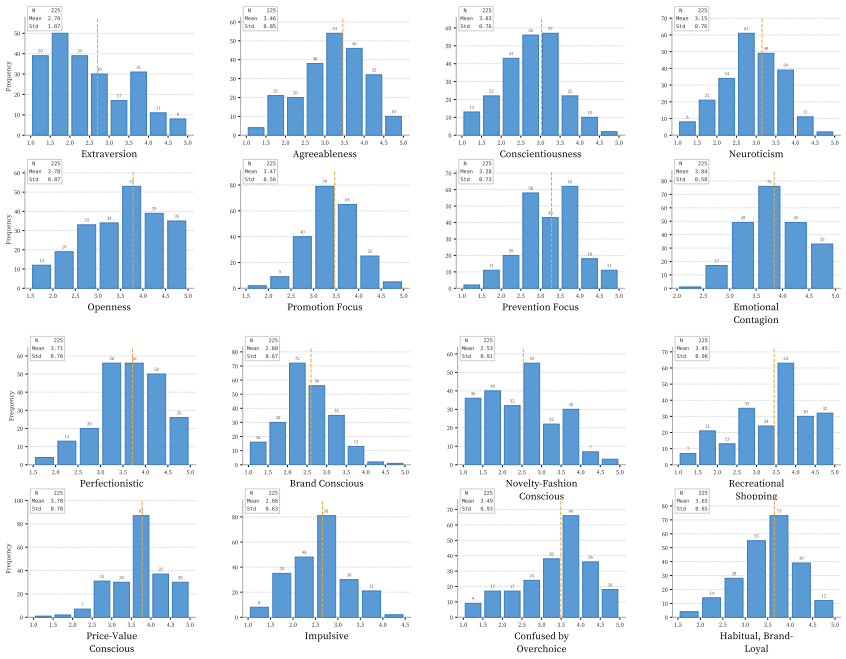


Figure 9: Distribution of participants personality and traits

D.2 Distribution of Annotation Items

We also report the overall distribution of participant ratings across all videos. Specifically, for each evaluation dimension, we first compute the Mean Opinion Score (MOS) for each video, and then summarize the distribution of these per-video MOS values across the full dataset, as shown in Figure 10– 14. This analysis provides an overview of the scoring range and distributional characteristics of the dataset under different subjective dimensions.

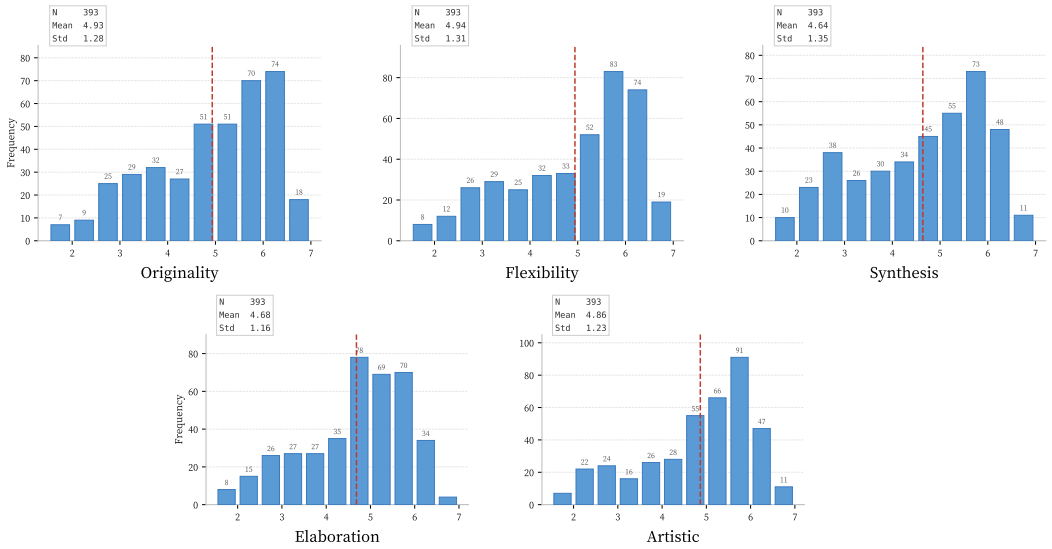


Figure 10: Distribution of per-video mean opinion scores (MOS) for creativity dimensions. Each subplot corresponds to one sub-dimension.

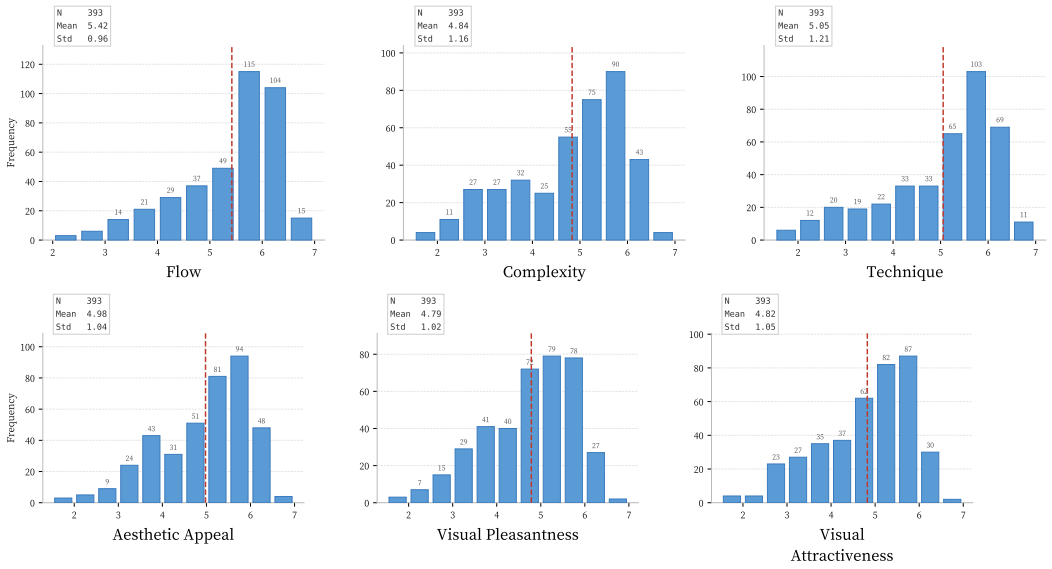


Figure 11: Distribution of per-video mean opinion scores (MOS) for Visual perceptual quality dimensions. Each subplot corresponds to one sub-dimension.

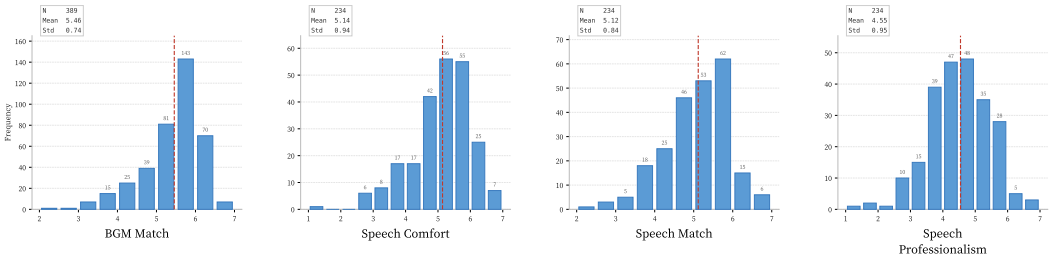


Figure 12: Distribution of per-video mean opinion scores (MOS) for Audio perceptual quality dimensions. Each subplot corresponds to one sub-dimension.

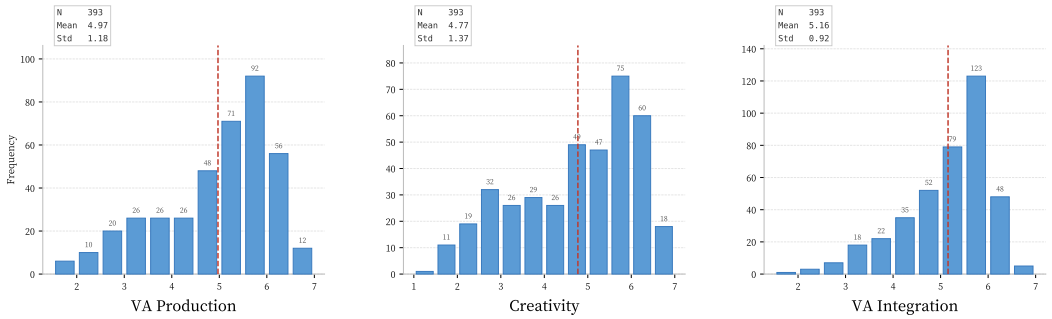


Figure 13: Distribution of per-video mean opinion scores (MOS) for General perceptual quality dimensions. Each subplot corresponds to one sub-dimension.

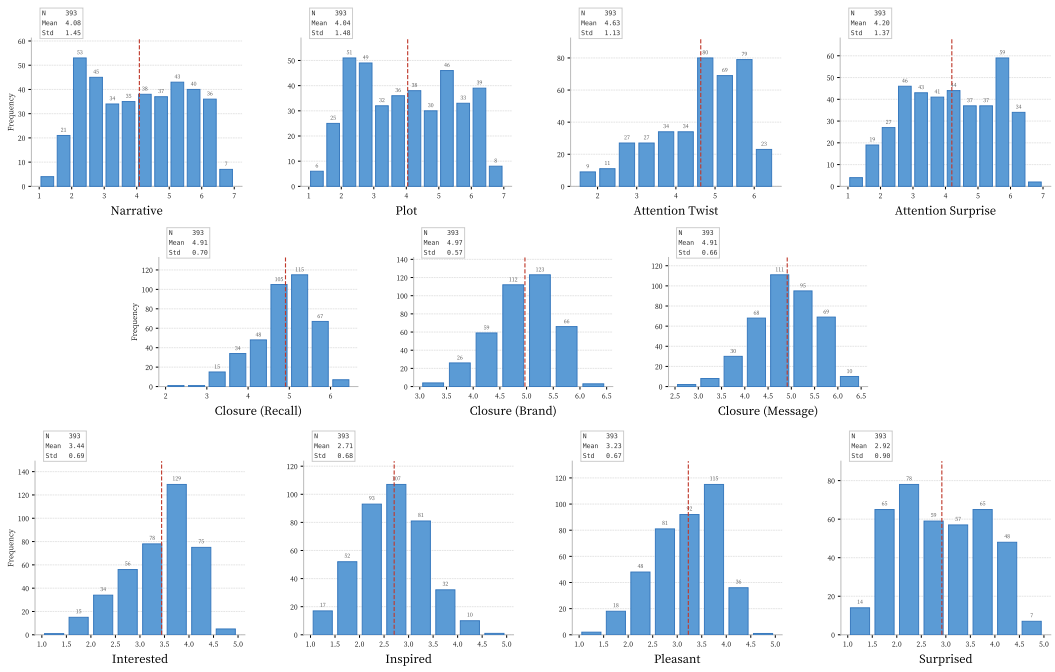


Figure 14: Distribution of per-video mean opinion scores (MOS) for information-processing annotation dimensions. Each subplot corresponds to one sub-dimension.

D.3 Open impressions from participants

Figure 15 shows the word clouds derived from the open-ended impression responses. We visualize positive and negative impressions separately. Overall, the 225 participants generated 6,783 impressions across 393 videos, averaging about 17 impressions per video. Among these, there were 5,050 positive impressions and 1,783 negative impressions. The high-frequency words indicate that participants often mentioned content-related aspects such as creativity, visuals, language, and music. At the same time, terms related to information processing were also frequently observed, including narrative, opening hook, contrast, suddenly, and interesting. These results suggest that the open-ended responses reflect not only overall evaluations of video content, but also the cognitive and perceptual cues that participants attended to during viewing.

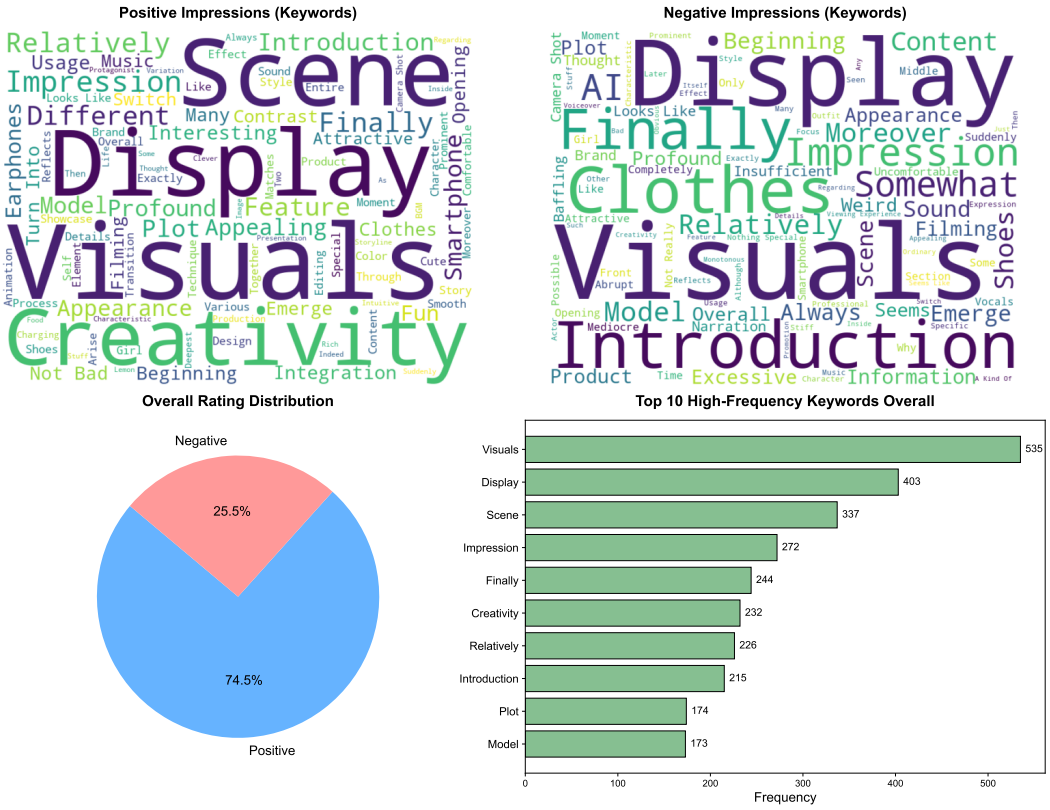


Figure 15: Wordclouds of raw human open impression points

E Additional Experiment results

E.1 Experiments Compute Resources

All models evaluated in this study were accessed through their official APIs and were not locally deployed. The environment required for analysis is described in the code documentation.

E.2 Evaluation Metrics

This section provides the definitions of the evaluation metrics used for rating tasks and explains their roles in analyzing MLLM–human alignment. Since our goal is not single-label prediction but alignment between model judgments and human subjective responses at multiple levels, we adopt a set of complementary metrics that capture trend-level correlation, rating-level agreement, and absolute deviation, details see Table 10.

Table 10: Evaluation metrics for rating-based task evaluation.

Metric	Alignment emphasis	Interpretation
Kendall's τ	Rank-order alignment	Captures whether MLLMs preserve the relative ordering of items similarly to humans, independent of exact score distances.
Pearson's r	Linear trend alignment	Captures whether MLLM and human ratings rise and fall together across items in a linearly consistent manner.
Linear-weighted Cohen's κ	Absolute rating agreement	Captures agreement on the original rating categories, with chance correction and heavier penalties for larger disagreements.
MAE	Absolute score deviation	Captures the average magnitude of score differences between MLLMs and humans on the original scale.

Table 11: Results on open impression extraction measured by weighted Recall@K and weighted Precision@K, grouped by evaluator pipeline.

Model	GPT-5.4				Gemini-3.5-Flash				DeepSeek-V4-Pro			
	WR@3	WP@3	WR@5	WP@5	WR@3	WP@3	WR@5	WP@5	WR@3	WP@3	WR@5	WP@5
Closed-source Models												
Gemini-2.5-Pro	0.4416	0.1199	0.5606	0.0892	0.4749	0.0969	0.6054	0.0742	0.4128	0.0967	0.5156	0.0752
Gemini-3-Flash	0.4528	0.1178	0.5625	0.0864	0.4692	0.0917	0.5905	0.0696	0.4175	0.0949	0.5085	0.0741
Gemini-3.1-Pro	0.4885	0.1271	0.5882	0.0914	0.5354	0.1043	0.6335	0.0753	0.4537	0.1018	0.5500	0.0790
GPT-5.4	0.5310	**0.1460**	**0.6700**	**0.1108**	0.5281	0.1026	0.6944	**0.0808**	0.4891	**0.1103**	0.6210	**0.0890**
Qwen3.5-Omni-Plus	0.5529	0.1392	0.6578	0.0993	**0.6251**	**0.1100**	**0.7343**	0.0782	**0.5241**	0.1096	**0.6358**	0.0831
Qwen3.5-Plus	0.5508	0.1382	0.6612	0.1009	0.5783	0.1053	0.7033	0.0775	0.4820	0.1012	0.5993	0.0813
Open-source Models												
Kimi-k2.5	0.5247	0.1262	0.6516	0.0944	0.5453	0.0962	0.6826	0.0728	0.5033	0.1068	0.6231	0.0831
Qwen3-VL-30B	0.4533	0.1037	0.5704	0.0781	0.4794	0.0734	0.6128	0.0556	0.4362	0.0820	0.5512	0.0654
Qwen3-VL-32B	0.4652	0.1092	0.6016	0.0848	0.4697	0.0733	0.6233	0.0577	0.4282	0.0813	0.5588	0.0666
Qwen3-VL-235B	0.4844	0.1165	0.6092	0.0879	0.5112	0.0860	0.6447	0.0652	0.4854	0.0964	0.6044	0.0761
Qwen3-Omni-Flash	0.4939	0.1134	0.6235	0.0862	0.5001	0.0786	0.6413	0.0618	0.4628	0.0858	0.5953	0.0690
Qwen3.5-27B	**0.5551**	0.1356	0.6646	0.0971	0.5758	0.1034	0.7029	0.0755	0.5091	0.1048	0.6182	0.0796
Qwen3.5-35B	0.5218	0.1248	0.6372	0.0919	0.5548	0.0937	0.6800	0.0692	0.4774	0.0960	0.6007	0.0763

E.3 Support-weighted Salient-Impression Evaluation

Each reference impression point has a support count, defined as the number of raw human responses mapped to it during reference-set construction. To account for how frequently each impression was mentioned, we report support-weighted Recall@K (WR@K) and Precision@K (WP@K) in Table 11, alongside the unweighted results in Table 3. The weighted metrics give greater credit to matches with higher-support reference points, while using the same rank-order matching and duplicate-credit rules as the unweighted metrics. Support-weighted Recall@K is generally higher than unweighted Recall@K, consistent with stronger coverage of widely shared impressions under support weighting. The two views are complementary: unweighted results retain equal emphasis on reference points regardless of their support.

E.4 MAE and Directional Scoring Bias

In the main results, we primarily report alignment metrics such as correlation and consistency. To complement these analyses, we additionally report mean absolute error (MAE) between MLLM and human ratings across all dimensions, as shown in Table 12. Absolute score deviations are larger on multimodal perceptual-quality dimensions than on creativity dimensions, mirroring the ranking-order contrast in Table 2, while signed biases are predominantly positive.

We further examine whether MLLMs systematically score videos higher or lower than humans by computing the mean signed difference between model and human ratings, defined as MLLM − Human. The results are shown in Table 13. Overall, most models show positive signed differences across dimensions, indicating a general tendency to overestimate relative to human ratings. This overestimation is especially pronounced for multimodal perceptual-quality dimensions, whereas the bias is comparatively smaller on creativity-related dimensions.

E.5 Audio Modality Screening

Because the evaluated models include both fully multimodal models and vision-language (VL) models without direct audio input, we first assess whether each model can correctly determine the presence of audio-related content in a video, specifically background music (BGM) and voice/narration.

Table 12: Performance evaluation of different models across Creativity and Multimodal Perception Quality. Best results are in bold; second-best results are underlined.

Metric	Model	Creativity						Multimodal Subjective Cognition											
		Divergence					Generally	Visual						Audio-Visual Match				Generally	
		Orig.	Flex.	Synth.	Elab.	Art.		Cmplx	Tech	Flow	Appeal	Plea	Attr	BGM	SMatch	SComf	SPro	VACoord	ProdQ
	Closed-source Models																		
MAE	Gemini-2.5-Pro	0.908	0.785	1.231	0.963	1.014	1.040	1.033	1.030	1.026	0.997	1.106	1.604	1.218	1.122	0.754	1.013	1.699	1.442
	Gemini-3-Flash	**0.541**	**0.609**	0.832	0.799	0.849	0.668	1.079	1.261	1.142	0.997	1.187	1.453	1.105	1.236	0.880	1.556	1.571	1.490
	Gemini-3.1-Pro	0.768	0.737	0.976	**0.733**	0.835	0.740	1.032	0.948	0.752	**0.847**	**0.925**	**1.086**	0.807	1.024	**0.581**	0.894	1.137	**1.116**
	GPT-5.4	0.639	0.766	0.876	0.777	0.970	0.831	1.105	1.015	**0.652**	1.166	1.147	1.191	0.477	1.000	0.650	**0.612**	0.811	1.180
	Qwen3.5-Omni-Plus	0.984	0.962	1.411	0.855	1.180	1.253	1.456	1.414	1.149	1.429	1.516	1.648	1.224	1.481	1.241	1.019	1.638	1.731
	Qwen3.5-Plus	0.623	0.677	0.755	0.813	0.901	**0.646**	1.143	1.219	0.731	1.346	1.421	1.317	0.550	0.857	0.693	0.853	1.051	1.446
	Open-source Models																		
	Kimi-k2.5	0.733	0.776	0.828	0.976	1.026	0.904	1.163	1.283	0.813	1.372	1.399	1.467	0.614	1.010	0.818	0.896	1.015	1.416
	Qwen3-VL-30B	0.774	0.930	0.820	1.024	1.032	1.085	1.195	1.268	0.889	1.397	1.464	1.477	0.652	1.037	0.867	1.226	1.113	1.513
	Qwen3-VL-32B	0.752	0.722	0.729	0.886	**0.828**	0.724	0.902	1.056	0.704	1.432	1.541	1.457	**0.456**	1.020	0.637	0.729	**0.752**	1.314
	Qwen3-VL-235B	0.863	0.839	0.829	0.877	0.952	0.855	**0.857**	**0.868**	0.680	1.103	1.241	1.236	0.465	**0.790**	0.695	0.831	0.911	1.131
	Qwen3-Omni-Flash	1.034	1.244	1.023	1.396	1.156	1.396	1.741	1.621	1.011	1.739	1.896	1.690	1.295	1.485	1.224	1.284	1.517	1.813
	Qwen3.5-27B	0.741	0.915	**0.711**	0.910	0.917	0.819	1.448	1.361	1.021	1.409	1.386	1.626	0.885	1.141	0.938	0.894	1.225	1.535
	Qwen3.5-35B	0.702	0.765	0.722	0.997	1.010	0.730	1.317	1.078	0.781	1.171	1.201	1.391	0.536	0.801	0.804	0.762	0.830	1.222

Table 13: MLLM scoring bias relative to human ratings across all dimensions. Values represent mean difference (MLLM − Human). Values represent mean difference (MLLM score − Human score). ↑ Red: MLLM overestimates; ↓ Blue: MLLM underestimates. Significance (FDR-corrected t-test) shown below each value: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$; unlabelled = $p \geq .05$. **Bold** = largest absolute bias per dimension; Underline = second largest.

Metric	Model	Creativity						Multimodal Subjective Cognition											
		Divergence					General	Visual						Audio-Visual Match				Generally	
		Orig.	Flex.	Synth.	Elab.	Art.		Cmplx	Tech	Flow	Appeal	Plea	Attr	BGM	SMatch	SComf	SPro	VACoord	ProdQ
	Closed-source Models																		
Scoring Bias	Gemini-2.5-Pro	↓-0.034	↑+0.014	↑+0.310 ***	↑+0.360 ***	↑+0.489 ***	↑+0.445 ***	↑+0.803 ***	↑+0.912 ***	↑+0.976 ***	↑+0.782 ***	↑+0.905 ***	↑+1.591 ***	↑+1.179 ***	↑+1.044 ***	↑+0.297 ***	↑+0.777 ***	↑**+1.658** ***	↑+1.370 ***
	Gemini-3-Flash	↑+0.086 *	↑+0.030	↑+0.333 ***	↑+0.417 ***	↑+0.183 ***	↑+0.463 ***	↑+1.041 ***	↑+1.247 ***	↑**+1.134** ***	↑+0.950 ***	↑+1.165 ***	↑+1.451 ***	↑+1.099 ***	↑+1.228 ***	↑+0.837 ***	↑**+1.556** ***	↑+1.571 ***	↑+1.487 ***
	Gemini-3.1-Pro	↓-0.310 ***	↓-0.325 ***	↓-0.348 ***	↓-0.306 ***	↓-0.191 ***	↓-0.053	↑+0.895 ***	↑+0.852 ***	↑+0.664 ***	↑+0.676 ***	↑+0.695 ***	↑+1.046 ***	↑+0.755 ***	↑+0.934 ***	↑+0.169 **	↑+0.637 ***	↑+1.098 ***	↑+1.029 ***
	GPT-5.4	↑+0.344 ***	↑+0.590 ***	↑+0.669 ***	↑+0.362 ***	↑+0.873 ***	↑+0.695 ***	↑+1.068 ***	↑+0.966 ***	↑+0.408 ***	↑+1.156 ***	↑+1.103 ***	↑+1.161 ***	↑+0.166 ***	↑+0.971 ***	↑+0.358 ***	↑+0.328 ***	↑+0.718 ***	↑+1.136 ***
	Qwen3.5-Omni-Plus	↑+0.844 ***	↑+0.714 ***	↑**+1.347** ***	↑+0.482 ***	↑+0.964 ***	↑+1.201 ***	↑+1.455 ***	↑+1.373 ***	↑+1.065 ***	↑+1.398 ***	↑+1.467 ***	↑+1.646 ***	↑+1.221 ***	↑**+1.473** ***	↑+1.184 ***	↑+0.792 ***	↑+1.632 ***	↑+1.714 ***
	Qwen3.5-Plus	↓-0.012	↓-0.098 *	↑+0.260 ***	↑+0.093	↑+0.542 ***	↑+0.327 ***	↑+1.090 ***	↑+1.121 ***	↑+0.600 ***	↑+1.283 ***	↑+1.384 ***	↑+1.306 ***	↑+0.420 ***	↑+0.799 ***	↑+0.525 ***	↑+0.586 ***	↑+1.013 ***	↑+1.403 ***
	Open-source Models																		
	Kimi-k2.5	↑+0.577 ***	↑+0.600 ***	↑+0.475 ***	↑+0.825 ***	↑+0.813 ***	↑+0.822 ***	↑+1.129 ***	↑+1.261 ***	↑+0.747 ***	↑+1.358 ***	↑+1.379 ***	↑+1.463 ***	↑+0.532 ***	↑+0.968 ***	↑+0.692 ***	↑+0.679 ***	↑+0.989 ***	↑+1.397 ***
	Qwen3-VL-30B	↑+0.266 ***	↑+0.673 ***	↑+0.453 ***	↑+0.769 ***	↑+0.592 ***	↑+0.913 ***	↑+1.158 ***	↑+1.242 ***	↑+0.823 ***	↑+1.385 ***	↑+1.444 ***	↑+1.467 ***	↑+0.558 ***	↑+1.012 ***	↑+0.786 ***	↑+1.168 ***	↑+0.992 ***	↑+1.506 ***
	Qwen3-VL-32B	↓-0.318 ***	↑+0.018	↓-0.279 ***	↑+0.174 **	↑+0.204 ***	↑+0.311 ***	↑+0.823 ***	↑+1.035 ***	↑+0.509 ***	↑+1.420 ***	↑+1.524 ***	↑+1.454 ***	↑+0.202 ***	↑+0.958 ***	↑+0.288 ***	↓-0.194 *	↑+0.271 ***	↑+1.305 ***
	Qwen3-VL-235B	↓-0.379 ***	↑+0.196 ***	↓-0.247 ***	↑+0.274 ***	↑+0.226 ***	↑+0.114 *	↑+0.642 ***	↑+0.670 ***	↑+0.366 ***	↑+1.035 ***	↑+1.186 ***	↑+1.213 ***	↑+0.182 ***	↑+0.428 ***	↑+0.016	↑+0.054	↑+0.017	↑+1.062 ***
	Qwen3-Omni-Flash	↑**+0.888** ***	↑**+1.189** ***	↑+0.872 ***	↑**+1.352** ***	↑**+1.040** ***	↑**+1.354** ***	↑**+1.734** ***	↑**+1.618** ***	↑+0.963 ***	↑**+1.736** ***	↑**+1.890** ***	↑**+1.690** ***	↑**+1.288** ***	↑+1.466 ***	↑**+1.194** ***	↑+1.227 ***	↑+1.512 ***	↑**+1.810** ***
	Qwen3.5-27B	↑+0.139 **	↑+0.597 ***	↓-0.031	↑+0.270 ***	↑+0.446 ***	↑+0.423 ***	↑+1.426 ***	↑+1.319 ***	↑+0.924 ***	↑+1.383 ***	↑+1.348 ***	↑+1.610 ***	↑+0.811 ***	↑+1.089 ***	↑+0.830 ***	↑+0.291 **	↑+1.186 ***	↑+1.512 ***
	Qwen3.5-35B	↑+0.031	↑+0.111 *	↓-0.076	↑+0.259 ***	↑+0.611 ***	↑+0.343 ***	↑+1.296 ***	↑+0.959 ***	↑+0.537 ***	↑+1.118 ***	↑+1.111 ***	↑+1.372 ***	↑+0.321 ***	↑+0.527 ***	↑+0.476 ***	↑+0.006	↑+0.639 ***	↑+1.134 ***

This screening step is important because model–human consistency on audio-related evaluation dimensions is only meaningful when both the model and human annotators agree that the relevant audio modality is present in the video. For human judgments, we use majority voting to determine the reference label for each video: a video is treated as containing BGM or voice/narration if more than half of the annotators indicate its presence. For model judgments, each model is asked the corresponding detection question three times per video; if at least two of the three responses indicate the presence of BGM or voice/narration, the model prediction is recorded as positive, and otherwise as negative. This procedure allows us to evaluate both the audio recognition ability of fully multimodal models and the audio inference ability of VL-only models. Only samples that are true negatives (TN) for all models are further included in the evaluation of the corresponding fine-grained dimensions, so that subsequent alignment analysis is conducted under consistent modality assumptions. The results are shown in Table 14. Fully multimodal models show clear advantages in audio presence screening, although their performance is still not perfect. At the same time, some VL-only models with relatively strong inference ability, such as GPT-5.4 and Qwen3.5-Plus, also achieve comparatively high detection performance.

E.6 Support Distribution of Reference Impressions

To characterize the reference sets constructed from open-ended human responses, Table 15 reports the number of consolidated impression points and the distribution of their support counts across the three evaluator pipelines. Support is the number of raw human impressions mapped to a reference

Table 14: Audio gate classification performance of each model on BGM and Voice detection. Acc. = Accuracy; Prec. = Precision; Spec. = Specificity. **Bold** = best per column; Underline = second best.

Model	BGM				Voice			
	Acc.	Prec.	Recall	F1	Acc.	Prec.	Recall	F1
Closed-source Models								
Gemini-2.5-Pro	**0.9771**	0.9791	0.9973	**0.9881**	0.9644	0.9367	**1.0000**	0.9673
Gemini-3-Flash	0.9618	0.9616	**1.0000**	0.9804	0.9695	0.9452	**1.0000**	0.9718
Gemini-3.1-Pro	0.9695	0.9789	0.9894	0.9841	**0.9822**	0.9673	**1.0000**	**0.9834**
GPT-5.4	0.8422	0.9564	0.8750	0.9139	0.8550	0.8319	0.9082	0.8684
Qwen3.5-Omni-Plus	0.9593	**0.9839**	0.9734	0.9786	0.9771	**0.9806**	0.9758	0.9782
Qwen3.5-Plus	0.9237	0.9577	0.9628	0.9602	0.8321	0.8190	0.8744	0.8458
Open-source Models								
Kimi-k2.5	0.9262	0.9554	0.9681	0.9617	0.8651	0.9477	0.7874	0.8601
Qwen3-VL-30B	0.8499	0.9648	0.8750	0.9177	0.6947	0.6408	0.9565	0.7675
Qwen3-VL-32B	0.5827	0.9818	0.5745	0.7249	0.7608	0.9185	0.5990	0.7251
Qwen3-VL-235B	0.6463	0.9684	0.6516	0.7790	0.8473	0.9106	0.7874	0.8445
Qwen3-Omni-Flash	0.9567	0.9567	**1.0000**	0.9779	0.8601	0.7901	**1.0000**	0.8827
Qwen3.5-27B	0.8550	0.9597	0.8856	0.9212	0.8270	0.9017	0.7536	0.8210
Qwen3.5-35B	0.7277	0.9622	0.7447	0.8396	0.7837	0.9420	0.6280	0.7536

Table 15: Number and support-count distribution of consolidated reference impressions across evaluator pipelines.

Evaluator pipeline	# Points	Points/video	Mean support	Median support	Support $= 1$ (%)	Support ≤ 2 (%)	Support ≤ 3 (%)
GPT-5.4	2635	6.70	3.72	3	16.81	36.05	52.10
Gemini-3.5-Flash	2574	6.55	2.94	2	27.58	50.58	68.80
DeepSeek-V4-Pro	2222	5.65	3.62	3	21.60	43.30	60.40

point. The final three columns give the percentages of retained reference points supported by exactly one, at most two, and at most three raw impressions, respectively.

The consolidated sets contain both recurrent and less frequently mentioned impressions. Mean support ranges from 2.94 to 3.72 raw impressions per reference point, while 16.81%–27.58% of retained points have support from a single raw impression. Thus, consolidation groups responses that express shared impressions without restricting the reference set to frequently mentioned views.

The consolidated sets contain both recurrent and less frequently mentioned impressions. Mean support ranges from 2.94 to 3.72 raw impressions per reference point, while 16.81%–27.58% of retained points have support from a single raw impression. Thus, consolidation groups responses that express shared impressions without restricting the reference set to frequently mentioned views.

F Process-Level Alignment under Progressive Exclusion by Rating Standard Deviation and Mean

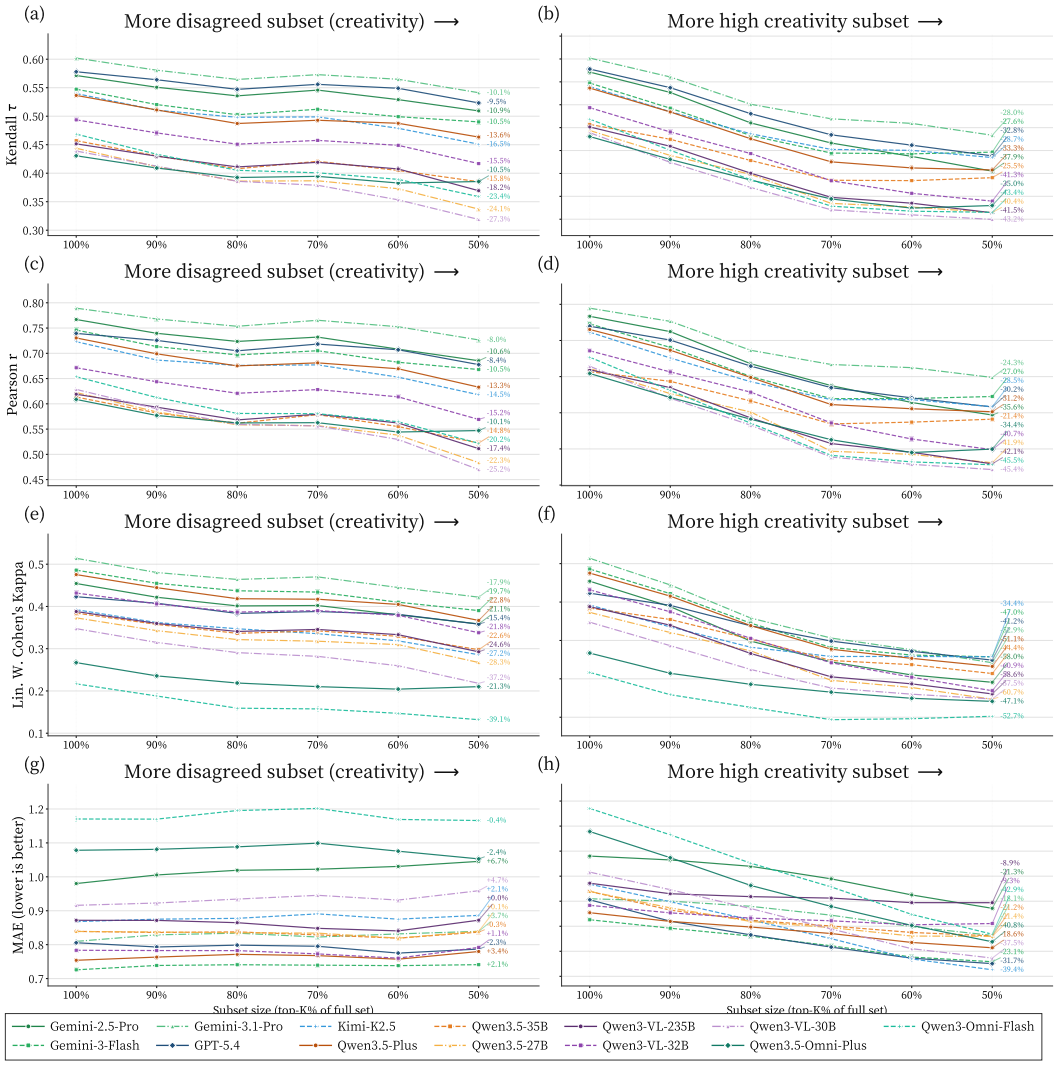


Figure 16: Change in MLLM–human alignment for Creativity under descending sample exclusion.

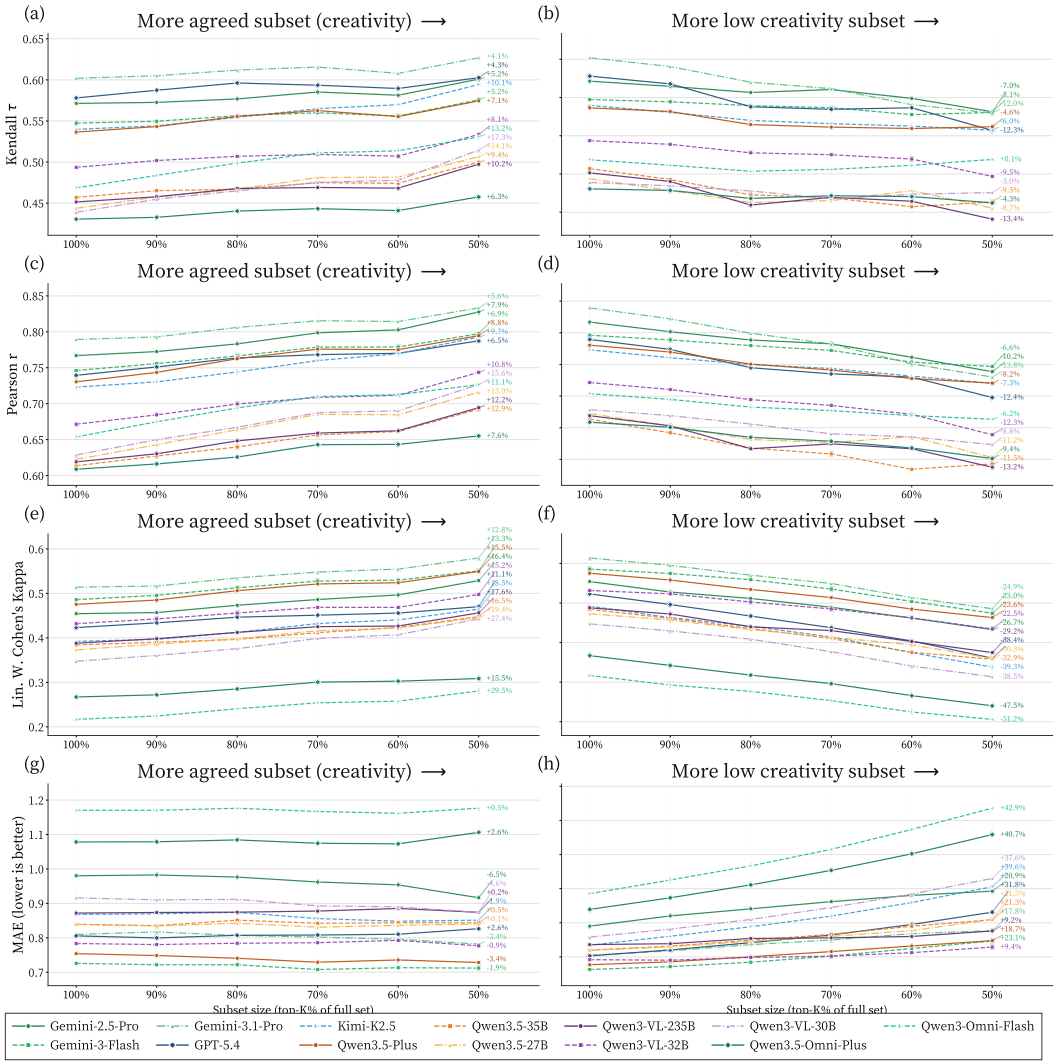


Figure 17: Change in MLLM–human alignment for Creativity under ascending sample exclusion.

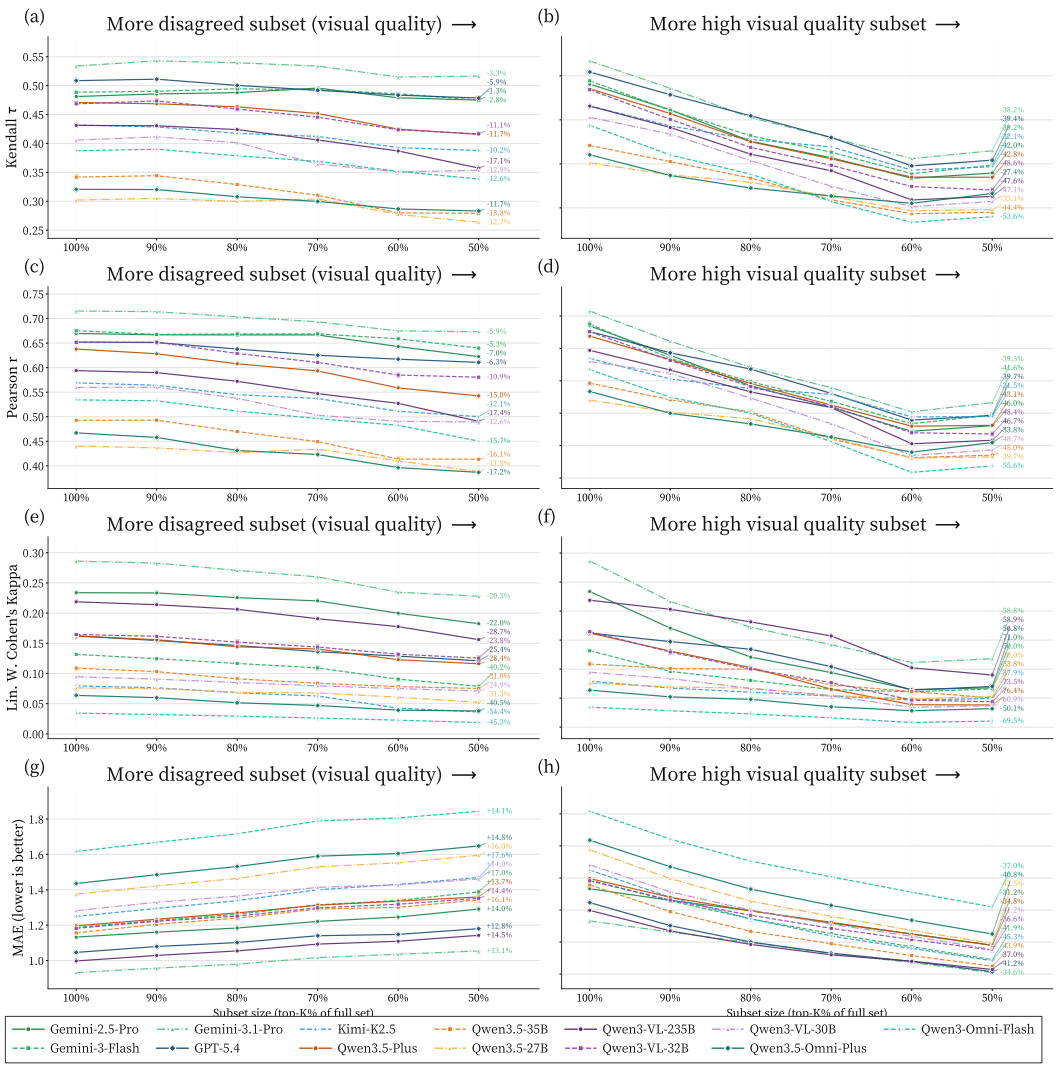


Figure 18: Change in MLLM–human alignment for Visual Quality under descending sample exclusion.

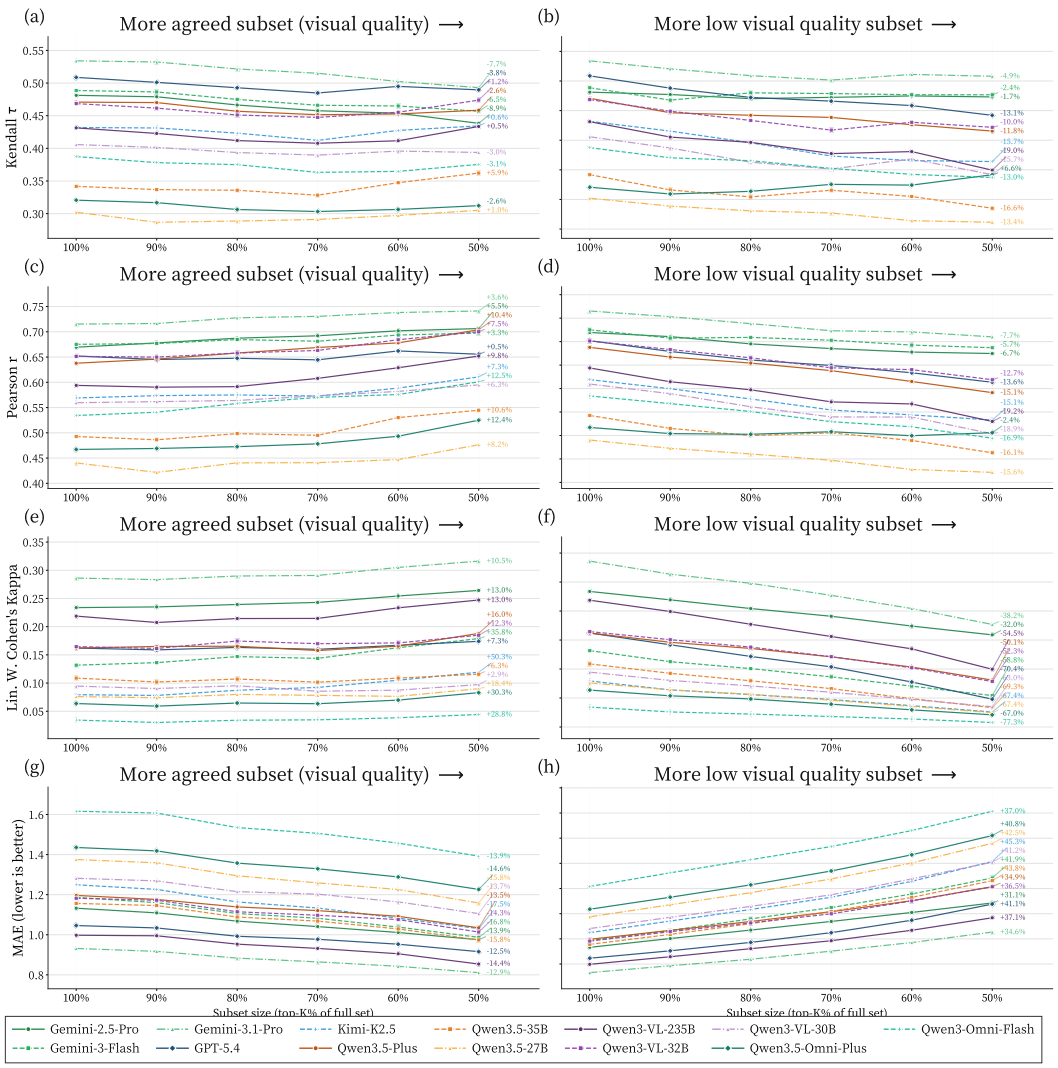


Figure 19: Change in MLLM–human alignment for Visual Quality under ascending sample exclusion.

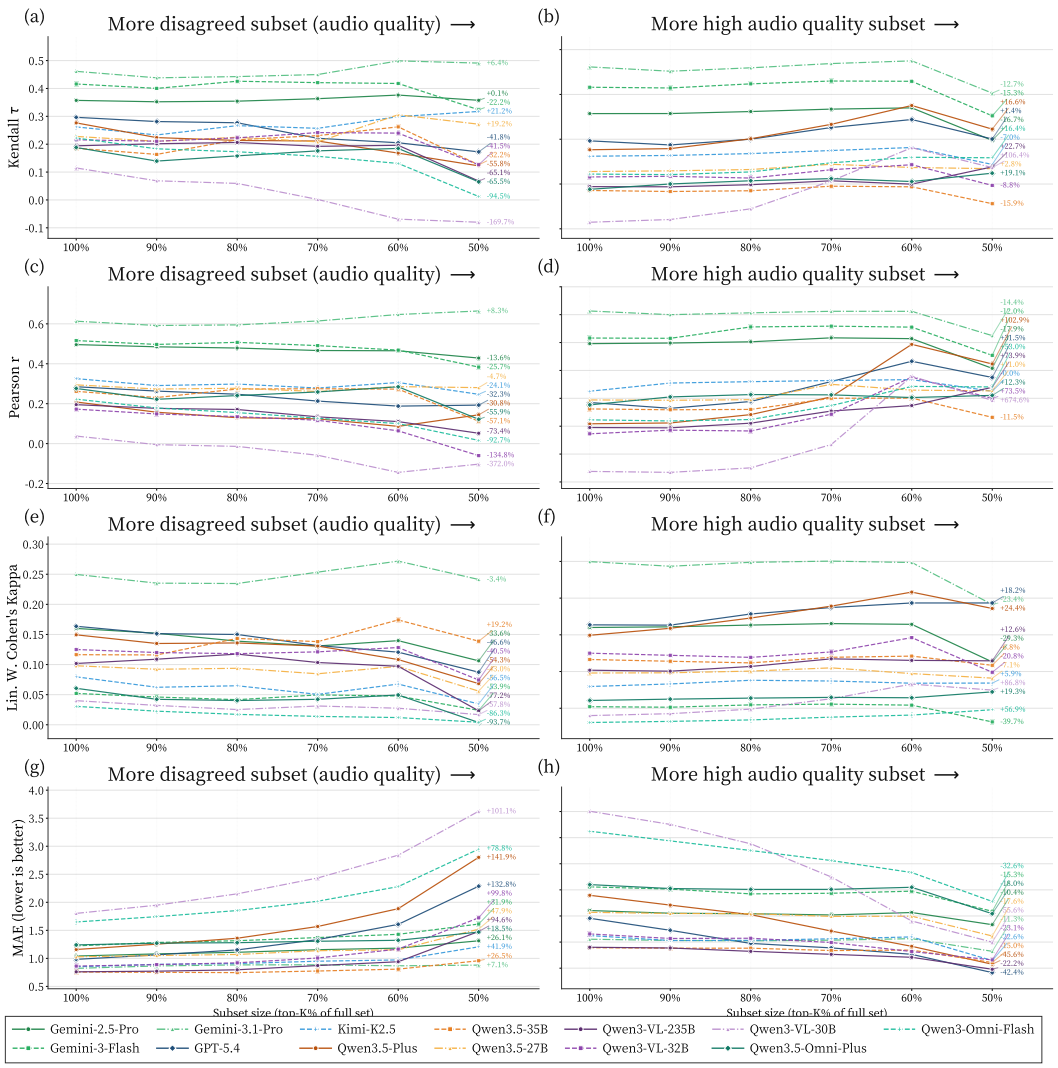


Figure 20: Change in MLLM–human alignment for Audio Quality under descending sample exclusion.

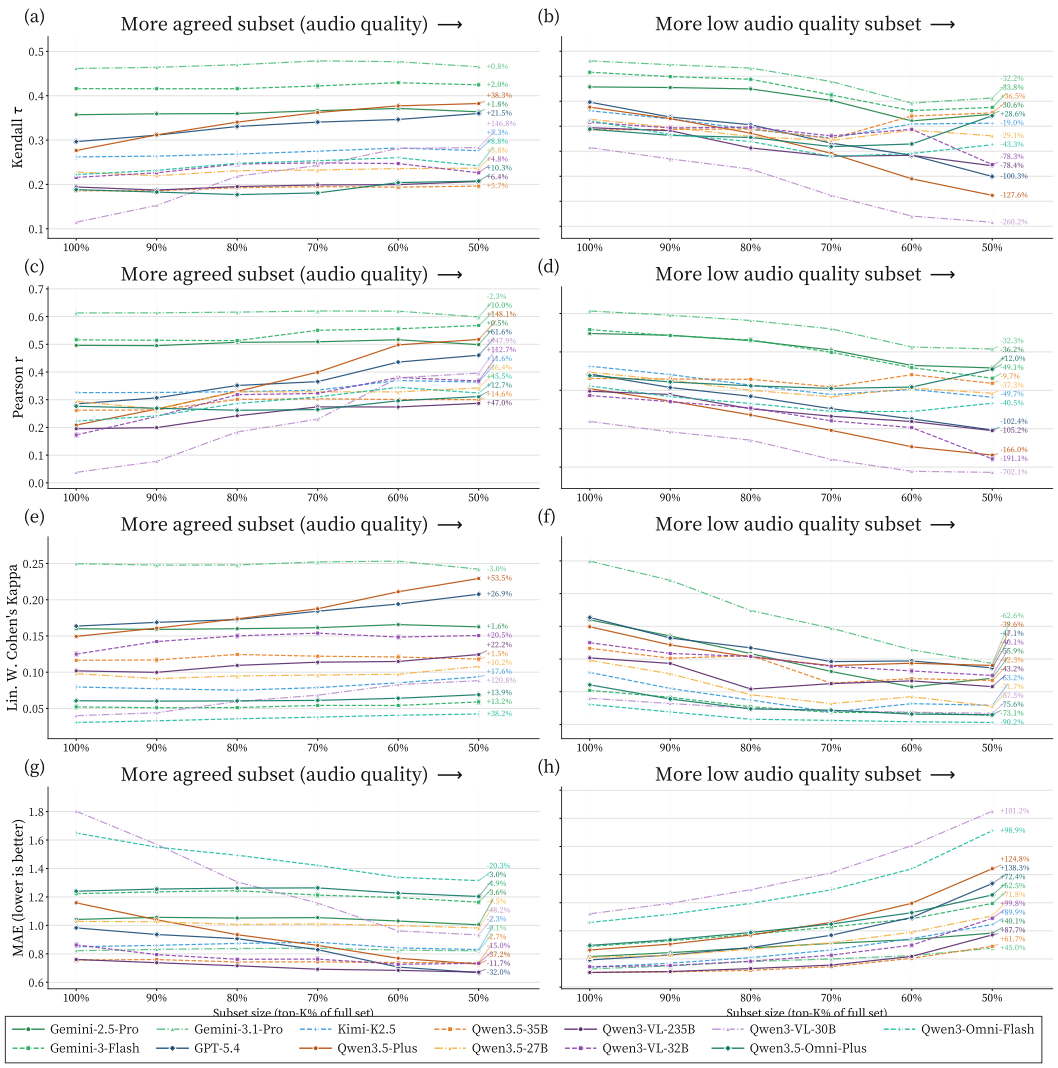


Figure 21: Change in MLLM–human alignment for Audio Quality under ascending sample exclusion.

G Full Trait-Based Alignment Results

The main text presents selected trait-based results for Creativity and Visual Perceptual Quality, focusing on the Big Five, regulatory focus, and emotional contagion. Figures 22–24 provide the full mixed-effects interaction results for all measured traits across Creativity, Visual Perceptual Quality, and Audio Perceptual Quality, respectively. Each figure shows the estimated interaction coefficients (γ_3) and their FDR-corrected significance, allowing trait-related patterns of model–human correspondence to be compared across models and evaluation dimensions.



Figure 22: Full mixed-effects interaction estimates (γ_3) for MLLM–human correspondence on **Creativity**. Significance is FDR-corrected: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.



Figure 23: Full mixed-effects interaction estimates (γ_3) for MLLM–human correspondence on **Visual Perceptual Quality**. Significance is FDR-corrected: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.



Figure 24: Full mixed-effects interaction estimates (γ_3) for MLLM–human correspondence on **Audio Perceptual Quality**. Significance is FDR-corrected: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

H SUPPLEMENTARY PROCESS-LEVEL ALIGNMENT THROUGH INFORMATION-PROCESSING DIMENSIONS

This section presents the complete supplementary results for the process-level analysis. Figure 25 provides a radar-chart overview of how model performance on multimodal perceptual-quality evaluation changes after introducing information-processing dimensions, summarized across four evaluation metrics. Tables 16 and 17 report the detailed numerical results for Creativity and Multimodal Perceptual Quality, respectively, when information-processing dimensions are incorporated. In addition, Figures 26 and 27 present the corresponding correlation and partial-correlation heatmaps for the creativity and perceptual-quality dimensions.

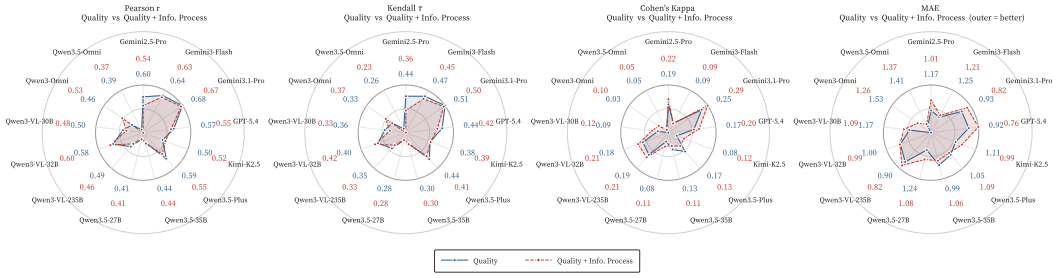


Figure 25: Performance change on multimodal perceptual quality evaluation after adding information-processing.

Table 16: Detailed process-level results for Creativity with information-processing dimensions.

Metric	Model	Creativity								Information Processing										
		Divergence					General			Narrative		Attn.Catch		DPC			Positive Affect			
		Orig.	Flex.	Synth.	Elab.	Art.	OvCr	VACoord	ProdQ	Sty	Plot	Hook	UnExp	DPC-1	DPC-2	DPC-3	Int.	Surp.	Insp.	Pleas.
Kendall's τ	**Closed-source Models**																			
	Gemini-2.5-Pro	0.424	0.418	0.436	0.274	0.304	0.429	0.277	0.339	0.586	0.600	0.331	0.436	0.321	-0.084	0.128	0.288	0.450	0.214	0.231
	Gemini-3-Flash	0.579	0.502	0.539	0.320	0.387	0.559	0.411	0.466	0.647	0.628	0.479	0.553	0.380	0.066	0.183	0.468	0.534	0.342	0.370
	Gemini-3.1-Pro	0.581	0.555	0.582	0.446	0.472	0.600	0.467	0.492	0.635	0.630	**0.481**	0.588	**0.422**	0.098	0.248	**0.511**	0.590	0.404	**0.449**
	GPT-5.4	0.621	**0.615**	**0.650**	0.420	**0.533**	0.632	**0.503**	0.497	**0.727**	**0.715**	0.338	0.615	0.298	-0.040	0.191	0.404	0.581	**0.424**	0.302
	Qwen3.5-Omni-Plus	**0.637**	0.399	0.527	0.336	0.322	0.625	0.341	0.373	0.657	0.652	0.327	0.577	0.319	0.112	**0.313**	0.278	0.558	0.251	0.186
	Qwen3.5-Plus	0.598	0.599	0.601	0.323	0.470	0.627	0.457	0.467	0.695	0.708	0.365	0.603	0.349	0.050	0.255	0.311	0.575	0.247	0.198
	Open-source Models																			
	Kimi-k2.5	0.617	0.608	0.628	0.408	0.461	**0.644**	0.471	**0.525**	0.684	0.673	0.428	**0.624**	0.373	0.035	0.204	0.404	**0.612**	0.272	0.387
	Qwen3-VL-30B	0.467	0.473	0.532	0.342	0.389	0.498	0.422	0.410	0.595	0.582	0.392	0.492	0.261	0.189	0.182	0.310	0.459	0.288	0.333
	Qwen3-VL-32B	0.582	0.564	0.610	0.404	0.498	0.587	0.468	0.483	0.665	0.642	0.470	0.580	0.326	0.182	0.227	0.371	0.557	0.265	0.379
	Qwen3-VL-235B	0.504	0.442	0.519	0.367	0.381	0.499	0.377	0.426	0.584	0.576	0.377	0.529	0.327	0.118	0.257	0.333	0.514	0.215	0.335
	Qwen3-Omni-Flash	0.514	0.498	0.556	**0.480**	0.507	0.527	0.436	0.455	0.634	0.653	0.427	0.530	0.365	**0.247**	0.079	0.426	0.548	0.238	0.366
	Qwen3.5-27B	0.532	0.476	0.566	0.372	0.410	0.537	0.335	0.387	0.655	0.653	0.399	0.585	0.301	0.138	0.266	0.313	0.571	0.182	0.279
	Qwen3.5-35B	0.506	0.467	0.551	0.308	0.381	0.550	0.369	0.365	0.642	0.648	0.319	0.545	0.286	0.119	0.214	0.208	0.521	0.242	0.198
Pear Corr.	**Closed-source Models**																			
	Gemini-2.5-Pro	0.557	0.596	0.577	0.417	0.476	0.563	0.427	0.514	0.719	0.720	0.534	0.576	0.386	-0.172	0.171	0.390	0.600	0.347	0.322
	Gemini-3-Flash	0.786	0.753	0.760	0.530	0.593	0.776	0.611	0.686	0.824	0.795	0.687	0.733	0.546	0.081	0.228	0.611	0.704	0.508	0.499
	Gemini-3.1-Pro	0.787	0.799	0.779	**0.700**	**0.711**	0.807	0.632	**0.735**	0.803	0.795	**0.704**	0.765	**0.592**	0.152	0.337	**0.682**	0.749	0.560	**0.621**
	GPT-5.4	**0.818**	0.791	**0.824**	0.537	0.711	0.818	**0.683**	0.668	**0.883**	**0.870**	0.473	0.793	0.358	-0.095	0.186	0.494	0.736	**0.580**	0.391
	Qwen3.5-Omni-Plus	0.781	0.610	0.708	0.458	0.494	0.780	0.500	0.544	0.832	0.832	0.485	0.746	0.464	0.168	**0.401**	0.358	0.716	0.398	0.264
	Qwen3.5-Plus	0.794	0.799	0.798	0.475	0.673	0.816	0.623	0.655	0.855	0.869	0.541	0.788	0.468	0.086	0.353	0.399	0.745	0.415	0.306
	Open-source Models																			
	Kimi-k2.5	0.816	**0.804**	0.810	0.610	0.666	**0.824**	0.674	0.717	0.859	0.853	0.636	**0.799**	0.498	0.040	0.276	0.530	**0.773**	0.436	0.509
	Qwen3-VL-30B	0.668	0.711	0.741	0.498	0.574	0.713	0.622	0.594	0.777	0.766	0.568	0.679	0.363	0.266	0.261	0.469	0.617	0.454	0.476
	Qwen3-VL-32B	0.764	0.767	0.791	0.552	0.685	0.782	0.665	0.669	0.833	0.821	0.669	0.762	0.482	0.264	0.307	0.501	0.722	0.460	0.515
	Qwen3-VL-235B	0.678	0.611	0.689	0.490	0.524	0.662	0.537	0.575	0.748	0.743	0.516	0.703	0.455	0.173	0.353	0.417	0.680	0.385	0.456
	Qwen3-Omni-Flash	0.730	0.726	0.763	0.647	0.702	0.744	0.631	0.637	0.807	0.826	0.620	0.705	0.528	**0.350**	0.122	0.582	0.708	0.420	0.488
	Qwen3.5-27B	0.737	0.706	0.766	0.512	0.574	0.740	0.485	0.558	0.821	0.820	0.574	0.765	0.422	0.203	0.384	0.440	0.740	0.343	0.379
	Qwen3.5-35B	0.693	0.663	0.747	0.406	0.519	0.741	0.526	0.509	0.815	0.831	0.473	0.724	0.407	0.155	0.291	0.288	0.676	0.383	0.268
Cohen's textbfKappa	**Closed-source Models**																			
	Gemini-2.5-Pro	0.347	0.363	0.321	0.269	0.266	0.362	0.021	0.134	0.381	0.350	0.169	0.338	0.168	-0.056	0.030	0.149	0.331	0.106	0.066
	Gemini-3-Flash	0.509	0.422	0.496	0.198	0.301	0.402	0.017	0.096	0.583	0.543	0.140	0.441	0.220	-0.008	0.128	0.155	0.408	0.277	0.018
	Gemini-3.1-Pro	0.520	**0.533**	0.514	**0.451**	**0.455**	**0.547**	0.205	**0.348**	0.501	0.479	**0.460**	0.467	0.249	0.007	0.187	**0.391**	0.424	0.221	**0.230**
	GPT-5.4	**0.580**	0.490	0.505	0.282	0.369	0.501	**0.264**	0.201	**0.698**	**0.647**	0.254	0.396	0.170	-0.058	0.009	0.016	0.362	**0.362**	0.016
	Qwen3.5-Omni-Plus	0.208	0.260	0.229	0.244	0.173	0.172	0.025	0.046	0.552	0.579	0.030	0.253	0.208	0.021	**0.240**	0.004	0.225	0.150	0.017
	Qwen3.5-Plus	0.572	0.516	0.455	0.219	0.318	0.504	0.148	0.123	0.605	0.611	0.172	0.534	0.276	-0.001	0.176	0.033	0.403	0.194	0.009
	Open-source Models																			
	Kimi-k2.5	0.438	0.421	0.495	0.214	0.296	0.395	0.162	0.132	0.651	0.646	0.162	0.400	0.270	-0.016	0.138	0.097	0.355	0.165	0.071
	Qwen3-VL-30B	0.410	0.427	0.502	0.233	0.335	0.358	0.257	0.126	0.583	0.546	0.289	0.466	0.257	0.013	0.063	0.051	0.325	0.226	0.078
	Qwen3-VL-32B	0.552	0.499	**0.531**	0.306	0.436	0.459	0.253	0.114	0.622	0.571	0.176	**0.551**	0.213	0.003	0.220	0.047	**0.485**	0.258	0.061
	Qwen3-VL-235B	0.465	0.393	0.458	0.299	0.300	0.393	0.258	0.150	0.500	0.508	0.344	0.496	0.273	0.004	0.214	0.130	0.468	0.223	0.069
	Qwen3-Omni-Flash	0.429	0.325	0.513	0.230	0.332	0.346	0.150	0.100	0.545	0.568	0.165	0.466	**0.292**	**0.056**	0.051	0.075	0.376	0.170	0.041
	Qwen3.5-27B	0.498	0.425	0.519	0.302	0.342	0.469	0.154	0.124	0.569	0.561	0.181	0.516	0.189	0.024	0.168	0.090	0.429	0.149	0.020
	Qwen3.5-35B	0.402	0.352	0.507	0.177	0.276	0.383	0.150	0.101	0.592	0.572	0.130	0.372	0.228	-0.002	0.140	0.008	0.184	0.098	0.004
MAE	**Closed-source Models**																			
	Gemini-2.5-Pro	1.249	1.080	1.427	1.171	1.165	1.237	1.379	1.309	1.365	1.438	1.222	1.415	1.006	1.712	1.653	0.757	0.944	1.253	0.931
	Gemini-3-Flash	0.668	0.709	0.884	0.918	0.920	0.878	1.526	1.288	0.703	0.812	1.233	0.915	0.565	**0.889**	0.689	0.653	0.650	0.615	1.223
	Gemini-3.1-Pro	0.745	0.712	0.938	0.742	0.788	0.764	0.831	**0.829**	1.021	1.030	0.790	0.989	0.896	1.287	0.782	**0.529**	0.747	0.815	**0.814**
	GPT-5.4	**0.556**	**0.564**	0.771	**0.652**	**0.738**	**0.658**	**0.565**	0.848	**0.518**	**0.530**	**0.741**	0.939	**0.462**	0.921	1.081	0.805	0.691	**0.472**	0.849
	Qwen3.5-Omni-Plus	1.162	0.991	1.430	0.884	1.194	1.324	1.336	1.566	0.735	0.712	1.745	1.328	0.620	0.957	0.643	0.938	0.941	0.835	1.139
	Qwen3.5-Plus	0.623	0.632	0.872	0.909	0.986	0.711	0.807	1.295	0.724	0.693	1.058	0.782	0.524	1.064	0.895	0.741	0.679	0.696	1.480
	Open-source Models																			
	Kimi-k2.5	0.807	0.781	0.893	1.058	0.988	0.964	0.843	1.245	0.649	0.674	1.214	1.106	0.530	0.963	0.699	0.721	0.814	0.689	0.899
	Qwen3-VL-30B	0.778	0.818	0.760	0.983	0.938	0.926	0.760	1.363	0.749	0.793	0.988	0.894	0.622	1.371	0.942	1.114	0.843	0.676	1.109
	Qwen3-VL-32B	0.656	0.692	0.782	0.820	0.784	0.802	0.779	1.405	0.697	0.769	1.269	**0.756**	0.685	1.858	**0.579**	0.828	**0.571**	0.595	1.001
	Qwen3-VL-235B	0.764	0.855	0.883	0.881	0.894	0.822	0.706	1.163	0.914	0.913	0.842	0.811	0.616	1.941	0.716	0.598	0.579	0.605	0.891
	Qwen3-Omni-Flash	0.827	0.901	0.730	1.054	0.871	1.021	1.030	1.407	0.750	0.738	1.188	0.845	0.631	1.168	0.826	0.946	0.714	0.718	1.193
	Qwen3.5-27B	0.698	0.809	0.764	0.865	0.886	0.806	0.883	1.305	0.759	0.802	1.179	0.772	0.819	1.244	0.866	0.672	0.654	0.613	1.063
	Qwen3.5-35B	0.792	0.862	**0.722**	0.985	0.939	0.917	0.903	1.222	0.689	0.717	1.333	1.034	0.653	1.351	0.808	1.152	1.210	0.961	1.537

Table 17: Detailed process-level results for Multimodal Perceptual Quality with information-processing dimensions.

Metric	Model	Multimodal Perceptual Quality													Information Processing									
		Visual						Audio-Visual				General			Narrative		Attn.Catch		DPC			Positive Affect		
		Cmplx	Tech	Flow	Appeal	VPlea	Attr	BGM	SMatch	SComf	SPro	OvCr	VACoord	ProdQ	Sty	Plot	Hook	UnExp	DPC-1	DPC-2	DPC-3	Int.	Surp.	Insp.	Pleas.
Kendall's τ	**Closed-source Models**																							
	Gemini-2.5-Pro	0.432	0.487	0.368	0.418	0.314	0.305	0.276	0.241	0.417	0.253	0.461	0.345	0.474	0.574	0.561	0.356	0.427	0.326	-0.075	0.118	0.348	0.433	0.237	0.339
	Gemini-3-Flash	0.583	**0.604**	0.319	0.436	0.445	0.411	0.396	0.346	0.423	**0.398**	0.661	0.485	**0.582**	0.700	0.677	0.483	0.609	**0.415**	0.115	0.307	0.484	0.602	0.372	0.437
	Gemini-3.1-Pro	0.532	0.593	**0.514**	0.514	0.478	0.481	**0.471**	**0.443**	**0.538**	0.369	0.636	**0.491**	0.553	0.700	0.692	**0.494**	0.609	0.409	0.062	**0.357**	**0.501**	**0.620**	**0.390**	**0.479**
	GPT-5.4	0.561	0.537	0.396	**0.559**	**0.493**	0.435	0.341	0.236	0.304	0.267	0.614	0.413	0.456	**0.719**	**0.706**	0.253	0.587	0.269	-0.043	0.160	0.381	0.553	0.325	0.321
	Qwen3.5-Omni-Plus	**0.615**	0.316	-0.016	0.294	0.271	0.253	0.282	0.029	0.129	0.158	0.648	0.197	0.289	0.676	0.683	0.246	0.596	0.264	0.138	0.260	0.216	0.581	0.264	0.192
	Qwen3.5-Plus	0.560	0.532	0.389	0.418	0.412	0.444	0.464	0.256	0.288	0.241	**0.664**	0.466	0.494	0.689	0.685	0.390	0.590	0.314	0.114	0.260	0.354	0.575	0.278	0.289
	Open-source Models																								
	Kimi-k2.5	0.525	0.522	0.418	0.435	0.351	0.431	0.423	0.178	0.271	0.114	0.657	0.477	0.494	0.705	0.698	0.429	**0.616**	0.362	-0.075	0.144	0.404	0.608	0.261	0.369
	Qwen3-VL-30B	0.465	0.440	0.348	0.367	0.351	0.419	0.245	0.167	0.199	0.206	0.504	0.352	0.356	0.588	0.577	0.331	0.503	0.116	0.077	0.183	0.256	0.457	0.298	0.335
	Qwen3-VL-32B	0.547	0.535	0.413	0.503	0.464	**0.503**	0.256	0.319	0.330	0.302	0.614	0.383	0.525	0.648	0.643	0.460	0.563	0.293	0.208	0.254	0.328	0.524	0.259	0.377
	Qwen3-VL-235B	0.431	0.417	0.395	0.390	0.385	0.415	0.242	0.142	0.170	0.132	0.505	0.395	0.448	0.595	0.578	0.349	0.526	0.334	0.065	0.230	0.379	0.524	0.260	0.353
	Qwen3-Omni-Flash	0.449	0.446	0.354	0.426	0.379	0.407	0.385	0.273	0.275	0.323	0.475	0.361	0.415	0.632	0.622	0.368	0.545	0.354	**0.222**	0.036	0.281	0.498	0.195	0.340
	Qwen3.5-27B	0.422	0.404	0.155	0.308	0.283	0.286	0.270	0.192	0.145	0.173	0.545	0.313	0.377	0.666	0.660	0.314	0.561	0.222	0.084	0.172	0.278	0.536	0.150	0.198
	Qwen3.5-35B	0.417	0.381	0.226	0.318	0.317	0.339	0.340	0.233	0.130	0.218	0.596	0.366	0.353	0.671	0.655	0.342	0.544	0.290	0.050	0.178	0.183	0.521	0.180	0.201
Pear Corr.	**Closed-source Models**																								
	Gemini-2.5-Pro	0.651	0.732	0.572	0.605	0.508	0.492	0.464	0.334	0.568	0.398	0.621	0.504	0.701	0.681	0.663	0.518	0.588	0.443	-0.145	0.142	0.464	0.600	0.368	0.482
	Gemini-3-Flash	0.742	**0.803**	0.525	0.616	0.623	0.602	0.573	0.470	0.570	**0.558**	0.829	**0.658**	**0.796**	**0.878**	0.857	0.645	0.790	0.535	0.166	0.416	0.580	0.774	0.545	0.561
	Gemini-3.1-Pro	0.711	0.793	**0.676**	0.694	0.656	0.697	**0.668**	**0.536**	**0.671**	0.495	0.828	0.656	0.761	0.869	0.866	**0.697**	0.787	**0.579**	0.105	**0.464**	**0.653**	**0.792**	**0.564**	**0.644**
	GPT-5.4	0.697	0.722	0.554	**0.722**	**0.669**	0.575	0.493	0.277	0.382	0.377	0.790	0.525	0.617	0.874	**0.867**	0.373	0.762	0.308	-0.077	0.193	0.466	0.715	0.485	0.392
	Qwen3.5-Omni-Plus	**0.754**	0.560	0.051	0.438	0.425	0.375	0.436	0.062	0.185	0.257	0.815	0.369	0.504	0.841	0.849	0.378	0.774	0.402	0.209	0.332	0.288	0.745	0.415	0.241
	Qwen3.5-Plus	0.702	0.713	0.535	0.580	0.580	0.602	0.629	0.350	0.387	0.296	0.834	0.602	0.670	0.859	0.860	0.545	0.778	0.387	0.167	0.366	0.430	0.747	0.434	0.394
	Open-source Models																								
	Kimi-k2.5	0.668	0.701	0.550	0.591	0.521	0.606	0.589	0.217	0.343	0.212	**0.838**	0.613	0.660	0.869	0.860	0.587	**0.791**	0.460	-0.080	0.190	0.494	0.771	0.418	0.476
	Qwen3-VL-30B	0.647	0.611	0.490	0.512	0.508	0.598	0.359	0.265	0.316	0.307	0.693	0.579	0.541	0.763	0.753	0.524	0.678	0.179	0.115	0.269	0.392	0.613	0.460	0.467
	Qwen3-VL-32B	0.747	0.754	0.604	0.667	0.637	**0.702**	0.387	0.435	0.488	0.449	0.813	0.566	0.717	0.819	0.812	0.674	0.741	0.449	0.286	0.347	0.462	0.691	0.441	0.505
	Qwen3-VL-235B	0.586	0.650	0.567	0.534	0.524	0.590	0.332	0.182	0.211	0.200	0.678	0.544	0.657	0.760	0.739	0.528	0.699	0.470	0.093	0.323	0.495	0.690	0.439	0.471
	Qwen3-Omni-Flash	0.631	0.627	0.510	0.595	0.558	0.579	0.564	0.386	0.355	0.455	0.676	0.550	0.581	0.802	0.800	0.556	0.728	0.530	**0.316**	0.044	0.424	0.653	0.372	0.487
	Qwen3.5-27B	0.574	0.596	0.280	0.440	0.400	0.395	0.409	0.290	0.261	0.324	0.716	0.449	0.535	0.839	0.847	0.428	0.741	0.286	0.094	0.247	0.364	0.715	0.309	0.285
	Qwen3.5-35B	0.550	0.576	0.363	0.477	0.482	0.484	0.448	0.310	0.209	0.325	0.764	0.528	0.540	0.848	0.840	0.488	0.717	0.386	0.061	0.247	0.252	0.690	0.307	0.300
Cohen's Kappa	**Closed-source Models**																								
	Gemini-2.5-Pro	**0.374**	0.325	0.103	**0.344**	0.275	0.128	0.110	0.106	0.331	0.217	0.380	0.026	0.255	0.327	0.297	0.144	0.353	0.215	-0.035	0.016	0.204	0.317	0.091	**0.167**
	Gemini-3-Flash	0.204	0.127	0.036	0.165	0.122	0.081	0.028	0.023	0.143	0.050	0.448	0.008	0.084	0.660	0.607	0.071	0.472	0.218	0.017	**0.281**	0.052	0.418	0.260	0.011
	Gemini-3.1-Pro	0.367	0.350	0.235	0.337	**0.333**	**0.287**	0.235	**0.225**	**0.403**	**0.330**	**0.555**	0.135	**0.296**	0.576	0.564	**0.360**	**0.508**	**0.309**	0.007	0.267	**0.242**	0.456	0.258	0.143
	GPT-5.4	0.207	0.262	**0.239**	0.209	0.253	0.107	0.282	0.079	0.202	0.172	0.487	0.201	0.203	**0.680**	**0.639**	0.114	0.458	0.119	-0.044	0.011	0.015	0.465	**0.275**	0.008
	Qwen3.5-Omni-Plus	0.073	0.104	0.004	0.093	0.099	0.012	0.024	0.014	0.008	0.141	0.190	0.004	0.035	0.521	0.577	0.005	0.269	0.143	0.041	0.167	0.014	0.292	0.137	0.008
	Qwen3.5-Plus	0.150	0.188	0.120	0.106	0.098	0.073	**0.292**	0.092	0.115	0.148	0.365	0.079	0.102	0.651	0.630	0.069	0.469	0.170	0.026	0.175	0.024	0.397	0.138	0.006
	Open-source Models																								
	Kimi-k2.5	0.152	0.105	0.107	0.106	0.107	0.065	0.224	0.149	0.158	0.098	0.321	0.134	0.084	0.658	0.638	0.071	0.400	0.200	-0.094	0.099	0.030	0.388	0.168	0.026
	Qwen3-VL-30B	0.219	0.114	0.122	0.103	0.118	0.089	0.141	0.105	0.132	0.096	0.304	0.170	0.079	0.498	0.490	0.117	0.456	0.057	0.029	0.103	0.031	0.336	0.221	0.049
	Qwen3-VL-32B	0.314	0.276	0.189	0.130	0.129	0.087	0.172	0.184	0.331	0.273	0.385	0.290	0.147	0.588	0.564	0.064	0.506	0.237	0.006	0.186	0.010	0.452	0.211	0.028
	Qwen3-VL-235B	0.371	**0.377**	0.236	0.231	0.182	0.195	0.119	0.085	0.055	0.085	0.449	**0.331**	0.287	0.551	0.529	0.291	0.477	0.261	-0.002	0.167	0.139	**0.492**	0.224	0.072
	Qwen3-Omni-Flash	0.080	0.066	0.095	0.056	0.050	0.043	0.111	0.139	0.179	0.238	0.216	0.052	0.060	0.576	0.575	0.051	0.487	0.254	**0.066**	0.049	0.017	0.365	0.131	0.015
	Qwen3.5-27B	0.126	0.153	0.058	0.136	0.148	0.054	0.143	0.091	0.110	0.104	0.343	0.110	0.098	0.629	0.611	0.040	0.456	0.173	0.005	0.147	0.021	0.422	0.147	0.026
	Qwen3.5-35B	0.079	0.145	0.062	0.107	0.129	0.034	0.147	0.143	0.055	0.142	0.332	0.115	0.134	0.611	0.621	0.025	0.302	0.190	0.024	0.142	0.003	0.164	0.047	0.004
MAE	**Closed-source Models**																								
	Gemini-2.5-Pro	0.887	0.971	0.946	0.814	0.929	1.248	0.999	0.948	0.690	0.970	1.206	1.522	1.166	1.527	1.675	1.254	1.419	0.867	1.605	1.725	0.656	1.002	1.288	**0.787**
	Gemini-3-Flash	1.023	1.314	1.159	1.014	1.177	1.331	1.121	1.204	0.841	1.290	0.833	1.571	1.483	0.593	0.708	1.404	0.895	0.562	0.909	0.622	0.783	0.633	0.646	1.337
	Gemini-3.1-Pro	0.907	0.910	0.717	**0.795**	**0.836**	**0.917**	0.682	0.782	**0.566**	0.736	0.717	1.020	1.008	0.855	0.867	**0.885**	0.901	0.632	1.055	0.675	**0.611**	0.691	0.783	0.936
	GPT-5.4	0.847	**0.770**	**0.492**	0.942	0.928	1.021	**0.447**	0.858	0.612	**0.672**	**0.639**	**0.664**	**0.901**	**0.509**	**0.538**	0.957	0.820	**0.467**	0.777	1.083	0.815	0.568	**0.518**	0.880
	Qwen3.5-Omni-Plus	1.449	1.428	1.200	1.387	1.460	1.689	1.125	1.350	1.098	0.929	1.314	1.635	1.703	0.852	0.785	2.006	1.290	0.728	0.840	0.812	0.968	0.829	0.797	1.124
	Qwen3.5-Plus	1.192	1.286	0.819	1.416	1.475	1.322	0.514	0.927	0.823	0.791	0.976	1.045	1.471	0.646	0.637	1.409	0.890	0.540	**0.720**	0.773	0.860	0.703	0.729	1.510
	Open-source Models																								
	Kimi-k2.5	1.112	1.211	0.754	1.229	1.259	1.322	0.516	0.809	0.686	0.823	1.081	0.896	1.318	0.619	0.647	1.396	1.071	0.508	0.771	0.731	0.725	0.749	0.707	0.929
	Qwen3-VL-30B	1.064	1.238	0.882	1.318	1.370	1.413	0.596	0.848	0.742	1.035	1.034	1.049	1.483	0.854	0.873	1.450	0.829	0.664	1.093	0.806	1.414	0.746	0.728	1.339
	Qwen3-VL-32B	0.885	1.025	0.731	1.335	1.395	1.455	0.494	0.948	0.636	0.786	0.968	0.812	1.398	0.683	0.753	1.603	**0.764**	0.579	1.454	**0.597**	1.045	0.592	0.621	1.151
	Qwen3-VL-235B	**0.782**	0.784	0.632	0.911	0.983	0.989	0.483	**0.772**	0.832	0.955	0.853	0.772	0.996	0.805	0.841	0.956	0.850	0.641	1.737	0.678	0.657	**0.563**	0.592	0.956
	Qwen3-Omni-Flash	1.572	1.468	0.902	1.510	1.671	1.569	0.925	0.935	0.841	0.808	1.151	1.344	1.603	0.697	0.713	1.641	0.788	0.545	0.956	0.743	1.236	0.662	0.803	1.435
	Qwen3.5-27B	1.336	1.187	0.974	1.212	1.131	1.423	0.691	0.946	0.865	0.862	0.992	1.022	1.347	0.638	0.661	1.650	0.869	0.704	0.992	0.831	0.811	0.660	0.635	1.209
	Qwen3.5-35B	1.361	1.175	0.876	1.256	1.217	1.502	0.601	0.887	0.860	0.736	1.051	1.051	1.256	0.677	0.664	1.833	1.199	0.677	0.896	0.757	1.691	1.326	1.130	1.913

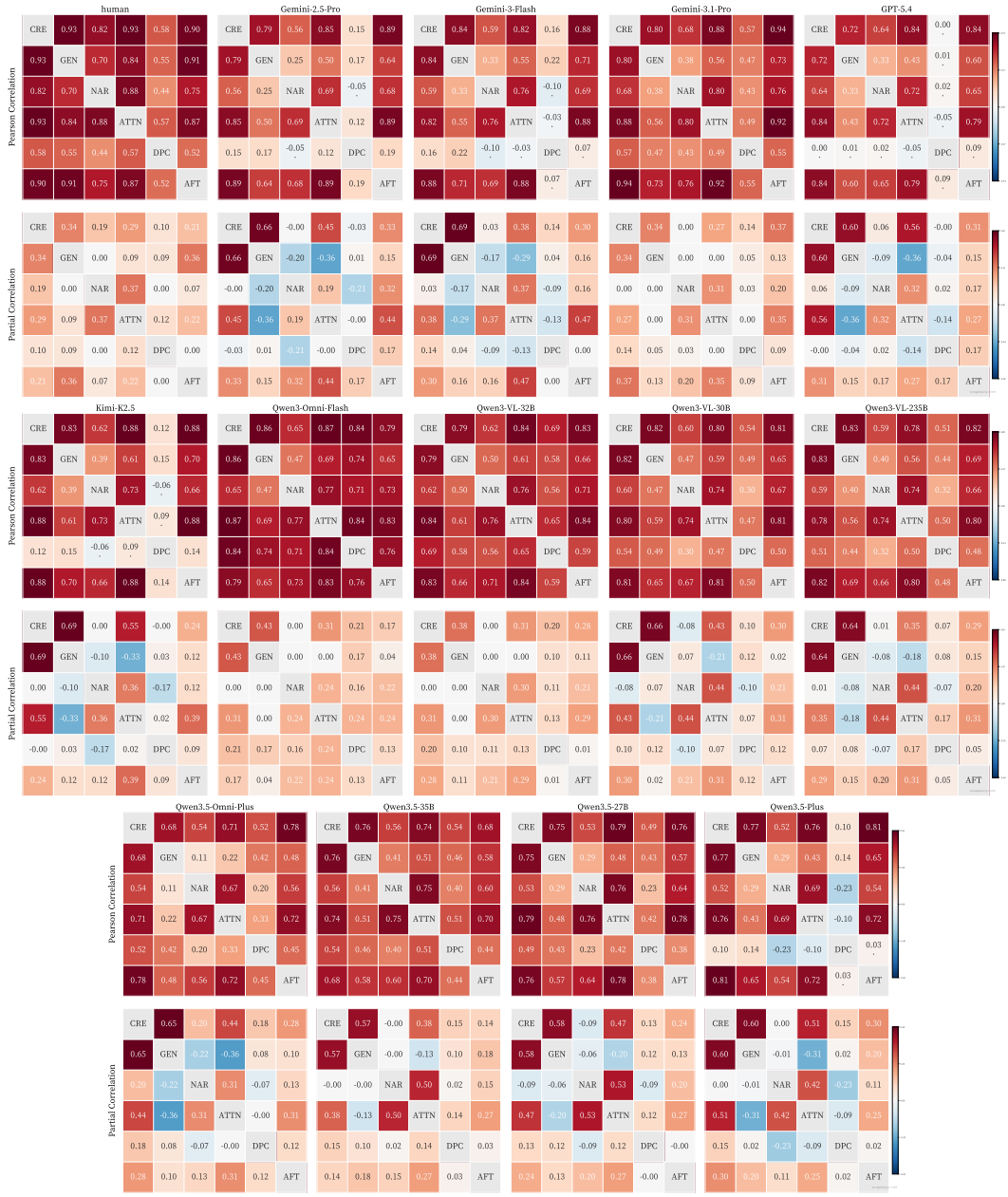


Figure 26: Correlation and partial-correlation heatmaps for Creativity and its associated information-processing dimensions. Cells with · are not significant at the 5% level.

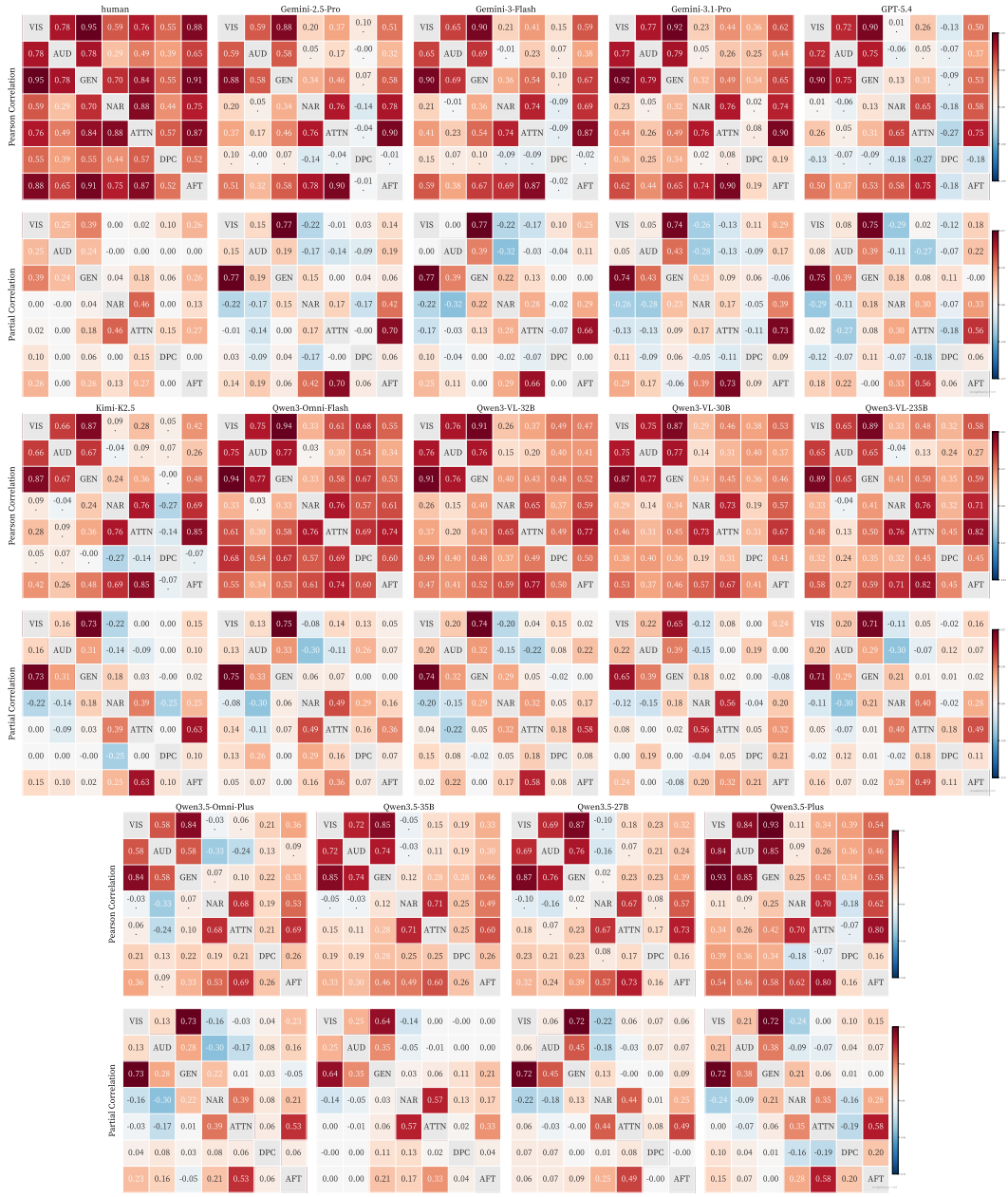


Figure 27: Correlation and partial-correlation heatmaps for Multimodal Perceptual Quality and its associated information-processing dimensions. Cells with · are not significant at the 5% level.

I PROMPTS FOR CONTENT AND INFORMATION PROCESSING EVALUATION

Prompt for Creativity Evaluation

Please simulate the genuine subjective impression of an ordinary viewer after watching an advertisement, and provide ratings on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree). There are no right or wrong answers; simply reflect your intuitive viewing experience. Output your response in standard JSON format, including the evaluation for each item only.
The evaluation items are as follows:
Creativity

- `q1`. This ad feels novel and distinct from ordinary ads.
- `q2`. This ad contains different creative ideas or elements.
- `q3`. This ad combines or connects elements that are normally unrelated.
- `q4`. This ad develops its basic ideas with numerous details, making them more intricate or sophisticated.
- `q5`. This ad has artistic appeal in its visual and/or verbal expression.
- `q6`. Overall, this ad is very creative.

Output requirements: Output only one valid JSON object, without any additional explanation or Markdown syntax.
Output format:

```
{
  "q1": {"score": 1-7, "reason": "your reason"},
  "q2": {"score": 1-7, "reason": "your reason"},
  "q3": {"score": 1-7, "reason": "your reason"},
  "q4": {"score": 1-7, "reason": "your reason"},
  "q5": {"score": 1-7, "reason": "your reason"},
  "q6": {"score": 1-7, "reason": "your reason"}
}
```

Please watch following video: {`video`}

Prompt for Creativity Evaluation

Please simulate the genuine subjective impression of an ordinary viewer after watching an advertisement, and provide ratings on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree). There are no right or wrong answers; simply reflect your intuitive viewing experience. Output your response in standard JSON format, including the evaluation for each item only.
The evaluation items are as follows:
Creativity

- `q1`. This ad feels novel and distinct from ordinary ads.
- `q2`. This ad contains different creative ideas or elements.
- `q3`. This ad combines or connects elements that are normally unrelated.
- `q4`. This ad develops its basic ideas with numerous details, making them more intricate or sophisticated.
- `q5`. This ad has artistic appeal in its visual and/or verbal expression.
- `q6`. Overall, this ad is very creative.

Output requirements:

- Output only one valid JSON object, without any additional explanation or Markdown syntax.

Output format:

```
{
  "q1": {"score": 1-7, "reason": "your reason"},
  "q2": {"score": 1-7, "reason": "your reason"},
  "q3": {"score": 1-7, "reason": "your reason"},
  "q4": {"score": 1-7, "reason": "your reason"},
  "q5": {"score": 1-7, "reason": "your reason"},
  "q6": {"score": 1-7, "reason": "your reason"},
}
```

Please watch following video: {`video`}

Prompt for Multimodal Perceptual Quality Evaluation

Please simulate the genuine subjective impression of an ordinary viewer after watching an advertisement, and provide ratings on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree). There are no right or wrong answers; simply reflect your intuitive viewing experience. Please ensure that your output is in standard JSON format, including the evaluation for each item.
The evaluation items are as follows:

Visual

- `v_q1`. The video contains rich visual elements in its composition.
- `v_q2`. The video demonstrates highly professional shooting techniques.
- `v_q3`. The video has a strong sense of visual flow in its editing and pacing.
- `v_q4`. The visuals of this video have high aesthetic appeal.
- `v_q5`. The visuals of this video have high visual pleasantness.
- `v_q6`. The visuals of this video have high visual attractiveness.

Audio

- `a_Q1`. Does this video contain background music (BGM)?
- `a_q1`. The background music matches the video content well.
- `a_Q2`. Does this video contain narration or spoken voice? (excluding sung vocals in the background music)
- `a_q2`. The narration or spoken voice is comfortable to listen to.
- `a_q3`. The narration or spoken voice matches the product and/or ad well.
- `a_q4`. The narration or spoken explanation sounds professional.

Overall

- `all_q1`. Overall, the visuals and sound of this video are well coordinated.
- `all_q2`. Overall, this video has high production quality.

Output format requirements:

- Output only one valid JSON object, without any additional explanation or Markdown syntax.
- For the two judgment items `a_Q1` and `a_Q2`, the values must be the strings `"Yes"` or `"No"`.
- For all other rating items, output an object containing `score` (an integer from 1 to 7) and `reason` (a string).
- If `a_Q1` is `"No"`, then both `score` and `reason` for `a_q1` must be set to `null`.
- If `a_Q2` is `"No"`, then both `score` and `reason` for `a_q2`, `a_q3`, and `a_q4` must all be set to `null`.

Example output format:

```
{
  "v_q1": {"score": 1-7, "reason": "your reason"},
  "v_q2": {"score": 1-7, "reason": "your reason"},
  "v_q3": {"score": 1-7, "reason": "your reason"},
  "v_q4": {"score": 1-7, "reason": "your reason"},
  "v_q5": {"score": 1-7, "reason": "your reason"},
  "v_q6": {"score": 1-7, "reason": "your reason"},
  "a_Q1": "No",
  "a_q1": {"score": null, "reason": null},
  "a_Q2": "Yes",
  "a_q2": {"score": 1-7, "reason": "your reason"},
  "a_q3": {"score": 1-7, "reason": "your reason"},
  "a_q4": {"score": 1-7, "reason": "your reason"},
  "all_q1": {"score": 1-7, "reason": "your reason"},
  "all_q2": {"score": 1-7, "reason": "your reason"}
}
```

Please watch following video: {`video`}

Prompt for Creativity & Information–processing Evaluation

Please simulate the genuine subjective impression of an ordinary viewer watching a short video advertisement on a mobile phone, and provide ratings on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree). For the emotion items, use a 5-point Likert scale (1 = not at all, 5 = very strongly). There are no right or wrong answers. Please respond entirely based on your intuitive viewing experience and genuine psychological reactions while watching. Please ensure that your output is in standard JSON format, including the evaluation for each question.
The evaluation items are as follows:

Creativity

- `c_q1`. This ad feels novel and distinct from ordinary ads.
- `c_q2`. This ad contains different creative ideas or elements.
- `c_q3`. This ad combines or connects elements that are normally unrelated.
- `c_q4`. This ad develops its basic ideas with numerous details, making them more intricate or sophisticated.
- `c_q5`. This ad has artistic appeal in its visual and/or verbal expression.

General

- `all_q1`. Overall, the visuals and sound of this video are well coordinated.
- `all_q2`. Overall, this video has high production quality.
- `all_q3`. Overall, this ad is very creative.

Narrative

- `Nar_q1`. I think this video has a strong sense of story.
- `Nar_q2`. I think the content of this video has a clear plot.

Attention Catch

- `attc_q1`. The opening of this video is very engaging and makes me want to keep watching.
- `attc_q2`. The content of this video felt unexpected to me.

Desire to Postpone Closure

- `dpc_q1`. While watching, I did not want to form a judgment about the brand/product too quickly, and instead wanted to see what additional information would follow.
- `dpc_q2`. I considered both the potential benefits and drawbacks of the brand/product rather than rushing to a judgment.
- `dpc_q3`. I felt that this short video did not provide enough information for me to make a reliable judgment.

Emotion (1 = not at all, 5 = very strongly)

- `emo_q1`. Interested
- `emo_q2`. Surprised
- `emo_q3`. Inspired
- `emo_q4`. Pleasant

Output format requirements:

- Output only one valid JSON object, without any additional explanation or Markdown syntax.
- For all rating items (`c_q`, `all_q`, `s_q`, `eng_q`, `th_q`, `emo_q`), output an object containing `score` (an integer from 1 to 7, or from 1 to 5 for emotion items) and `reason` (a brief string explaining the reason).

Example output format:

```
{
  "c_q1": {"score": 1-7, "reason": "your reason"},
  "c_q2": {"score": 1-7, "reason": "your reason"},
  "c_q3": {"score": 1-7, "reason": "your reason"},
  "c_q4": {"score": 1-7, "reason": "your reason"},
  "c_q5": {"score": 1-7, "reason": "your reason"},
  "all_q1": {"score": 1-7, "reason": "your reason"},
  "all_q2": {"score": 1-7, "reason": "your reason"},
  "all_q3": {"score": 1-7, "reason": "your reason"},
  "nar_q1": {"score": 1-7, "reason": "your reason"},
```

```
  "nar_q2": {"score": 1-7, "reason": "your reason"},
  "attc_q1": {"score": 1-7, "reason": "your reason"},
  "attc_q2": {"score": 1-7, "reason": "your reason"},
  "dpc_q1": {"score": 1-7, "reason": "your reason"},
  "dpc_q2": {"score": 1-7, "reason": "your reason"},
  "dpc_q3": {"score": 1-7, "reason": "your reason"},
  "emo_q1": {"score": 1-5, "reason": "your reason"},
  "emo_q2": {"score": 1-5, "reason": "your reason"},
  "emo_q3": {"score": 1-5, "reason": "your reason"},
  "emo_q4": {"score": 1-5, "reason": "your reason"}
}
```

Please watch following video: {video}

Prompt for Multimodal Perceptional Quality & Information Processing Evaluation

Please simulate the genuine subjective impression of an ordinary viewer watching a short video advertisement on a mobile phone, and provide ratings on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree). For the emotion items, use a 5-point Likert scale (1 = not at all, 5 = very strongly). There are no right or wrong answers. Please respond entirely based on your intuitive viewing experience and genuine psychological reactions while watching. Please ensure that your output is in standard JSON format, including the evaluation for each question.
The evaluation items are as follows:

Visual

- `v_q1`. The video contains rich visual elements in its composition.
- `v_q2`. The video demonstrates highly professional shooting techniques.
- `v_q3`. The video has a strong sense of visual flow in its editing and pacing.
- `v_q4`. The visuals of this video have high aesthetic appeal.
- `v_q5`. The visuals of this video have high visual pleasantness.
- `v_q6`. The visuals of this video have high visual attractiveness.

Audio

- `a_Q1`. Does this video contain background music (BGM)?
- `a_q1`. The background music matches the video content well.
- `a_Q2`. Does this video contain narration or spoken voice? (excluding sung vocals in the background music)
- `a_q2`. The narration or spoken voice is comfortable to listen to.
- `a_q3`. The narration or spoken voice matches the product and/or ad well.
- `a_q4`. The narration or spoken explanation sounds professional.

General

- `all_q1`. Overall, the visuals and sound of this video are well coordinated.
- `all_q2`. Overall, this video has high production quality.
- `all_q3`. Overall, this ad is very creative.

Narrative

- `Nar_q1`. I think this video has a strong sense of story.
- `Nar_q2`. I think the content of this video has a clear plot.

Attention Catch

- `attc_q1`. The opening of this video is very engaging and makes me want to keep watching.
- `attc_q2`. The content of this video felt unexpected to me.

Desire to Postpone Closure

- `dpc_q1`. While watching, I did not want to form a judgment about the brand/product too quickly, and instead wanted to see what additional information would follow.
- `dpc_q2`. I considered both the potential benefits and drawbacks of the brand/product rather than rushing to a judgment.
- `dpc_q3`. I felt that this short video did not provide enough information for me to make a reliable judgment.

Emotion (1 = not at all, 5 = very strongly)

- `emo_q1`. Interested
- `emo_q2`. Surprised
- `emo_q3`. Inspired
- `emo_q4`. Pleasant

Output format requirements:

- Output only one valid JSON object, without any additional explanation or Markdown syntax.
- For the two judgment items `a_Q1` and `a_Q2`, the values must be the strings `"Yes"` or `"No"`.
- For all other rating items (`v_q`, `a_q`, `all_q`, `s_q`, `eng_q`, `th_q`, `emo_q`), output an object containing `score` (an integer from 1 to 7, or from 1 to 5 for emotion items) and `reason` (a brief string).

- If a_Q1 is "No", then both score and reason for a_q1 must be set to null.
- If a_Q2 is "No", then both score and reason for a_q2, a_q3, and a_q4 must all be set to null.

Example output format:

```
{
  "v_q1": {"score": 1-7, "reason": "your reason"},
  "v_q2": {"score": 1-7, "reason": "your reason"},
  "v_q3": {"score": 1-7, "reason": "your reason"},
  "v_q4": {"score": 1-7, "reason": "your reason"},
  "v_q5": {"score": 1-7, "reason": "your reason"},
  "v_q6": {"score": 1-7, "reason": "your reason"},
  "a_Q1": "No",
  "a_q1": {"score": null, "reason": null},
  "a_Q2": "Yes",
  "a_q2": {"score": 1-7, "reason": "your reason"},
  "a_q3": {"score": 1-7, "reason": "your reason"},
  "a_q4": {"score": 1-7, "reason": "your reason"},
  "all_q1": {"score": 1-7, "reason": "your reason"},
  "all_q2": {"score": 1-7, "reason": "your reason"},
  "all_q3": {"score": 1-7, "reason": "your reason"},
  "nar_q1": {"score": 1-7, "reason": "your reason"},
  "nar_q2": {"score": 1-7, "reason": "your reason"},
  "attc_q1": {"score": 1-7, "reason": "your reason"},
  "attc_q2": {"score": 1-7, "reason": "your reason"},
  "dpc_q1": {"score": 1-7, "reason": "your reason"},
  "dpc_q2": {"score": 1-7, "reason": "your reason"},
  "dpc_q3": {"score": 1-7, "reason": "your reason"},
  "emo_q1": {"score": 1-5, "reason": "your reason"},
  "emo_q2": {"score": 1-5, "reason": "your reason"},
  "emo_q3": {"score": 1-5, "reason": "your reason"},
  "emo_q4": {"score": 1-5, "reason": "your reason"}
}
```

Please watch following video: {video}

Prompt for Consolidating Open-Ended Human Impressions into Video-Level Gold Points

You are a research assistant. Your task is to consolidate multiple human participants' open-ended responses about the same video into a gold-standard set of impression points for that video.

Task Objective

Given multiple participants' open-ended responses to the same video, please complete the following:

- Split multiple impression points within a response. A single response may contain several different impression points. First, split each response into more atomic points based on meaning.
- Merge semantically duplicate or highly similar points. Different participants may describe the same impression point using different wording. Merge such semantically similar points into one unified gold point.
- Retain important negative responses and lack-of-impression responses. Content such as "nothing particularly stood out," "rather ordinary," "the product was not clearly explained," "a bit confusing," "strong AI-generated feeling," or "none" should also be treated as valid impression points. Do not remove them, and pay attention to whether the impression is positive or negative.
- Distinguish the point itself from evaluative tendency. The same impression point may receive different evaluative tendencies from different participants. Therefore, separately organize the content of the impression point (`point`) and the evaluative tendency toward that point (`valence`).
- Construct video-level gold points. The final output should be a set of gold impression points at the video level, rather than a response-by-response restatement of the original human answers.

Merging Principles

- A gold point should express one relatively independent and identifiable impression point whenever possible.
- If one response mentions multiple ideas, split them apart.
- If multiple responses express the same core impression, merge them.
- Do not forcibly merge clearly different points.
- Do not mix broad summaries with specific points, unless that summary itself is repeatedly mentioned by multiple participants.
- For overly brief or vague responses (e.g., "authentic," "food," "refrigerator"), prioritize interpreting their most likely intended impression based on the response itself and the context of other responses to the same video.
- If a vague response can be reliably expanded, clarify its meaning in `description`.
- If a vague response cannot be reliably interpreted, retain it as a higher-level vague point instead of making unsupported guesses, and mark `"too_vague"` in `notes`.
- One original response may support multiple gold points.
- The same point may receive both positive and negative evaluations. In such cases, preserve both rather than forcing a single unified judgment.

Allowed Values for `point_type`

- `"scene_or_event"`: a specific scene, action, visual moment, or event
- `"plot_or_narrative"`: plot design, story progression, reversal, or narrative logic
- `"product_or_feature"`: the product itself, product functions, selling points, or design features
- `"audio_or_music"`: background music, sound, voice-over, or sound effects
- `"visual_style"`: visual style, color tone, composition, sense of technology, artistic look, or visual presentation
- `"editing_or_direction"`: editing, pacing, camera orchestration, transitions, or presentation technique
- `"character_or_performer"`: actors, roles, character image, or performance
- `"copy_or_dialogue"`: lines, slogan, copywriting, or verbal expression itself
- `"concept_or_theme"`: overall creative concept, thematic idea, or stylistic concept
- `"overall_impression"`: overall, summary-level, or relatively vague general impression
- `"lack_or_problem"`: lack of highlights, unclear expression, aversion, incongruity, strong AI-generated feeling, insufficient information, or other negative issues

Output Requirements
Please output a JSON object containing the following fields:
Top-level fields

- `video_summary`: a one-sentence summary of the main structure of human impressions of this video
- `gold_points`: an array in which each element represents one gold point

Each gold point must contain the following fields

- `point_id`: e.g., `"P1"`, `"P2"`
- `normalized_point`: a normalized expression of the impression point, written as one concise and complete sentence
- `point_type`: choose the single most appropriate category from the given list
- `description`: a more detailed explanation of the point, which may be slightly longer than `normalized_point`
- `support_count`: the number of responses that mention this point
- `support_response_ids`: a list of original response IDs supporting this point
- `representative_quotes`: 1–3 of the most representative original excerpts
- `valence_distribution`: statistics of positive/negative evaluations among the supporting responses, including `positive` and `negative`
- `majority_valence`: choose from `"positive"`, `"negative"`, or `"split"`
- `is_vague`: `true` or `false`
- `notes`: optional; used to explain ambiguity, potential confusion, or to mark `"too_vague"`

Additional Requirements

- The number of gold points should generally be controlled within 4–8, unless the human responses for this video are genuinely very rich.
- Do not split clearly repetitive points into multiple ones.
- Do not forcibly merge points that are actually different.
- If the same point clearly receives both positive and negative attitudes, preserve this split in `valence_distribution`.
- If a point is only a vague word but cannot be reliably expanded, do not delete it; retain it and mark `is_vague=true`.
- Do not output the intermediate splitting process or any explanatory text. Output only the final JSON.

Input Format
I will provide a response list. Each response contains:

- `response_id`
- `valence_label` (takes the value `"positive"` or `"negative"`)
- `content` (the participant's specific impression of the video)

Example input:

```
{
  "video_id": "xxx",
  "responses": [
    {
      "response_id": "R1",
      "valence_label": "positive",
      "content": "The music cuts off abruptly when the robot vacuum
          first appears."
    },
    {
      "response_id": "R2",
      "valence_label": "positive",
      "content": "It is very interesting. When the robot vacuum appears
          , it creates the amusing feeling that the audience guessed
          the wrong protagonist."
    }
  ]
}
```

Example output format:

```
{
  "video_summary": "The most prominent human impressions concentrate
      on the contrastive turn created by the product's appearance, the
       anthropomorphic visual setup, the musical atmosphere, and a
      small amount of negative feedback about insufficient product
      expression.",
  "gold_points": [
    {
      "point_id": "P1",
      "normalized_point": "The sudden appearance of the product
          interrupts the previously built-up atmosphere and creates a
          clear contrastive turn.",
      "point_type": "plot_or_narrative",
      "description": "The first half of the video gradually builds
          atmosphere or situational tension, and then the product
          suddenly appears and shifts the rhythm. This contrastive turn
           forms the main memorable point.",
      "support_count": 2,
      "support_response_ids": ["R1", "R2"],
      "representative_quotes": [
        "The music cuts off abruptly when the robot vacuum first appears
            .",
        "It is very interesting. When the robot vacuum appears, it
            creates the amusing feeling that the audience guessed the
            wrong protagonist."
      ],
      "valence_distribution": {
        "positive": 2,
        "negative": 0
      },
      "majority_valence": "positive",
      "is_vague": false,
      "notes": "The core of this point is the contrastive entrance and
          rhythmic disruption, rather than simply the product appearing
          ."
    }
  ]
}
```

Please analyze the following data: {content}

Prompt for Extracting the Top Five Memorable Impression Points

Please extract the most memorable impression points of this video from the perspective of what an ordinary viewer would actually notice, remember, and be willing to mention. Stay as close as possible to first-layer impressions formed under natural viewing conditions. Do not over-interpret, and do not invent meanings that are not clearly conveyed in the video. Prioritize answers based on content that can be directly observed in the video.
Please watch this video and answer:
What are the most memorable impressions this video leaves on you?
Please output the **five impression points that viewers are most likely to notice, remember, and mention**, and rank them by **degree of memorability**. Here, an impression point refers to a unit of content that is especially easy to notice, remember, evaluate, or bring up after watching. It may include, but is not limited to, the following:

- specific scenes, actions, visual moments, or events
- plot design, story progression, reversals, or narrative logic
- the product itself, product functions, selling points, or design features
- background music, sound, voice-over, or sound effects
- visual style, color tone, composition, sense of technology, aesthetic quality, or visual presentation
- editing, pacing, transitions, camera orchestration, or presentation style
- actors, roles, character image, or performance
- lines, slogans, copywriting, or verbal expression itself
- the overall creative concept, thematic idea, or stylistic concept
- obvious shortcomings, incongruity, unclear expression, aversive elements, or other problems

Please follow these requirements:

- Output exactly five impression points—no more and no fewer. The `rank` values must be 1, 2, 3, 4, and 5, with no repetition.
- Each impression point should express only one core idea as much as possible. Do not pack multiple different points into one sentence.
- Different impression points should be as independent and distinguishable as possible. Do not let two or more points actually refer to the same core content using different wording or emphasizing different aspects of it.
- Each impression point should be as specific as possible and correspond to a clear element in the video, such as a scene, expression style, line, product feature, or overall problem.
- Subjective words such as "interesting," "good-looking," "ordinary," or "off-putting" may be used, but specify as clearly as possible what exactly they refer to.
- If an impression point is negative, still output it as usual.
- Do not provide only vague summaries, and do not simply repeat overall judgments such as "this ad is good/bad."
- Do not write long analytical paragraphs, and do not output any extra explanatory text.
- The `evidence` field should be as concise as possible. Only state the most direct basis, and do not expand it into a long analysis.
- If the video itself does not contain many vivid and independently extractable memory points, do not force overly fine-grained splitting just to reach a more concrete level. Higher-level overall impressions may be output, but you should still explain as much as possible which specific contents, expressive choices, or overall features mainly give rise to that impression.

For each impression point, please provide the following information:

- `rank`: ranked by degree of memorability, where `1` is the most memorable and `5` is relatively less important
- `point`: one concise, specific, and distinguishable description of the impression point
- `evidence`: why this impression point is valid, i.e., what content, visuals, sounds, lines, or expressive basis in the video supports it
- `valence`: your own clear evaluative tendency toward this impression point; must be one of `positive` or `negative`
- `likely_human_valence_pattern`: your estimate of the overall attitude pattern that most people may have toward this impression point; must be one of `mostly_positive`, `mostly_negative`, or `mixed`

Specifically:

- If you think most people would likely view this point positively, use `mostly_positive`.
- If you think most people would likely view this point negatively, use `mostly_negative`.
- If you think this point is likely to provoke disagreement and different viewers may react in opposite ways, use `mixed`.

Please output the result in JSON using the following format:

```
{
  "impressions": [
    {
      "rank": 1,
      "point": "one concise impression point description",
      "evidence": "the corresponding content, visual, sound, line, or
          expressive basis in the video",
      "valence": "positive",
      "likely_human_valence_pattern": "mostly_positive"
    },
    {
      "rank": 2,
      "point": "one concise impression point description",
      "evidence": "the corresponding content, visual, sound, line, or
          expressive basis in the video",
      "valence": "negative",
      "likely_human_valence_pattern": "mixed"
    },
    {
      "rank": 3,
      "point": "one concise impression point description",
      "evidence": "the corresponding content, visual, sound, line, or
          expressive basis in the video",
      "valence": "positive",
      "likely_human_valence_pattern": "mostly_positive"
    },
    {
      "rank": 4,
      "point": "one concise impression point description",
      "evidence": "the corresponding content, visual, sound, line, or
          expressive basis in the video",
      "valence": "positive",
      "likely_human_valence_pattern": "mostly_positive"
    },
    {
      "rank": 5,
      "point": "one concise impression point description",
      "evidence": "the corresponding content, visual, sound, line, or
          expressive basis in the video",
      "valence": "negative",
      "likely_human_valence_pattern": "mostly_negative"
    }
  ]
}
```

Please watch following video: {`video`}

Prompt for Matching MLLM Impression Points to Human Gold Impression Points

You are an expert in advertising and subjective video understanding evaluation. Your task is to judge whether an impression point proposed by an AI after watching a video matches a particular human gold impression point, where the gold point has been consolidated from multiple human open-ended responses. You will be given:

- the full set of human gold impression points for the video
- one AI impression point to be judged

Your task is to:

- find the single best-matched gold point among all human gold points for the AI point
- return `null` if no gold point meets the matching standard
- output a matching score of `1`, `0.5`, or `0`

Scoring Rules

1 = Match

You may output `1` only if the AI point satisfies all of the following conditions:

- it expresses the same core impression as a gold point
- it contains all key information of that gold point, or expresses the same complete meaning using different wording or synonyms
- the AI's `likely_human_valence_pattern` is fully consistent with that gold point's `majority_valence`
- it does not contain conflicting, opposite, or unreasonable content

0.5 = Partial Match

You may output `0.5` only if the AI point satisfies all of the following conditions:

- it expresses the same core impression direction as a gold point
- it contains most, but not all, of the key information of that gold point
- the AI's `likely_human_valence_pattern` is not completely inconsistent with that gold point's `majority_valence`
- it does not contain conflicting, opposite, or unreasonable content

0 = No Match

Output `0` if any of the following applies:

- it does not contain any key information from any gold point
- it is not the same core impression point
- it only belongs to the same broad category, but not the same specific impression point
- the AI's `likely_human_valence_pattern` is fully inconsistent with the gold point's `majority_valence`
- it contains conflicting, opposite, or unreasonable content

Additional Rules

- **Valence consistency rule:** the correspondence between the AI point's `likely_human_valence_pattern` and the human gold point's `majority_valence` is as follows:
  - *Fully consistent:*
    - `mostly_positive` – `positive`
    - `mostly_negative` – `negative`
    - `mixed` – `split`
  - *Relatively consistent:*
    - `mostly_positive` – `split`
    - `mostly_negative` – `split`
    - `mixed` – `positive`
    - `mixed` – `negative`
  - *Fully inconsistent:*
    - `mostly_positive` – `negative`
    - `mostly_negative` – `positive`
- For the same AI point, at most one gold point may be assigned a score of `1`.

- If the AI point is related to multiple gold points, choose the single best-matched one as the final correspondence.
- Do not judge two points as matched merely because they belong to the same broad category such as music, visuals, function, creativity, atmosphere, or plot.
- If uncertain, judge strictly and prefer `0`.
- If the final score is `0`, this means the AI point does not match any human gold point. In this case, you must output:
  - `best_matched_gold_point_id`: `null`
  - `best_matched_gold_point`: `null`
  - `human_gold_majority_valence`: `null`

  Do not force a closest gold point when score = `0`. Only fill in the corresponding gold point when the final score is `1` or `0.5`.

Examples:{`examples`}
Input

- Human Gold Impression Points: {`human_gold_points`}
- AI Impression Point: {`ai_point`}

Output Requirements
Please output JSON only, without any additional text.
The format is as follows:

```
{
  "video_id": "xxx",
  "matches": [
    {
      "ai_rank": 1,
      "ai_point": "...",
      "ai_point_valence": "...",
      "human_gold_majority_valence": "...",
      "best_matched_gold_point_id": "P2",
      "best_matched_gold_point": "...",
      "score": 1,
      "reason": "..."
    },
    {
      "ai_rank": 2,
      "ai_point": "...",
      "ai_point_valence": "...",
      "human_gold_majority_valence": "...",
      "best_matched_gold_point_id": "P1",
      "best_matched_gold_point": "...",
      "score": 0.5,
      "reason": "..."
    },
    {
      "ai_rank": 3,
      "ai_point": "...",
      "ai_point_valence": "...",
      "human_gold_majority_valence": null,
      "best_matched_gold_point_id": null,
      "best_matched_gold_point": null,
      "score": 0,
      "reason": "..."
    },
    ...
  ]
}
```

Please analyze the following data: {`content`}