Measuring What Matters Beyond Text: Evaluating Multimodal Summaries by Quality, Alignment, and Diversity

Abid Ali* **and Diego Mollá-Aliod and Usman Naseem**
School of Computing, Macquarie University, Sydney, Australia
abidmeeraj@gmail.com, diego.molla-aliod@mq.edu.au, usman.naseem@mq.edu.au

Abstract

Multimodal Large Language Models (MLLMs) have facilitated Multimodal Summarization with Multimodal Output (MSMO), wherein systems generate concise textual summaries accompanied by salient visuals from multimodal sources. However, current MSMO evaluation remains fragmented: text quality, image–text alignment, and visual diversity are typically assessed in isolation using unimodal metrics, making it difficult to capture whether the modalities jointly support a faithful and useful summary. To address this gap, we introduce `MM-Eval`, a unified evaluation framework that integrates assessments of textual quality, cross-modal alignment, and visual diversity. `MM-Eval` comprises three components: (1) text quality, measured using OpenFActScore for factual consistency and G-Eval for coherence, fluency, and relevance; (2) image–text relevance, evaluated via an MLLM-as-a-judge approach; and (3) image-set diversity, quantified using Truncated CLIP Entropy. We calibrate `MM-Eval` through a learned aggregation model trained on the *mLLM-EVAL* news benchmark, aligning component contributions with human preferences. Our analysis reveals a text-dominant hierarchy in this setting, where factual consistency acts as a critical determinant of perceived overall quality, while visual relevance and diversity provide complementary signals. `MM-Eval` improves over heuristic aggregation baselines and provides an interpretable, reference-weak framework for comparative evaluation of multimodal summaries.

1 Introduction

Modern information consumption increasingly combines text with visuals, motivating Multimodal Summarization with Multimodal Output (MSMO): the task of generating textual summaries paired with relevant images that enrich and contextualize the content (Zhu et al., 2018, 2020). MSMO aims to produce coherent multimodal outputs that balance informativeness with cognitive load (Jiang et al., 2023), improving user satisfaction through faster gist comprehension via images while preserving detail in text (Zhu et al., 2018).

Despite its promise, MSMO has historically been constrained by the semantic gap: the challenge of aligning low-level visual signals with high-level textual semantics. Recent advances in Multimodal Large Language Models (MLLMs), such as GPT-4V, LLaVA (Liu et al., 2023a), and Qwen-VL (Wang et al., 2024), have helped bridge this gap by supporting joint encoding and reasoning across modalities (Chang et al., 2024). However, as generative capabilities improve, a critical bottleneck has emerged: the lack of robust, scalable, and reliable automatic evaluation methods for multimodal outputs (Zhuang et al., 2024).

Despite the inherently multimodal nature of MSMO, current evaluation protocols remain fragmented: a limitation we refer to as the "Silo Effect". Outputs are decomposed and assessed via unimodal metrics, each blind to cross-modal coherence. On the textual side, ROUGE remains dominant. While historically impactful, ROUGE is limited in its ability to capture semantic equivalence or factual consistency (Schluter, 2017; Zhang et al., 2024). Consequently, summaries that superficially match reference texts but contain hallucinated facts can still score highly (Lage and Ostermann, 2025).

On the visual side, Image Precision (IP) evaluates whether the model selects the exact images chosen by annotators, penalizing semantically valid alternatives, and ignoring alignment with textual content (see Section 2 for a detailed discussion).

Most critically, current metrics fail to assess the interaction between modalities: a system may score well on ROUGE and image selection independently while producing disconnected multimodal outputs. Figure 1 illustrates these limitations with toy examples.

*Corresponding Author

Findings of the Association for Computational Linguistics: ACL 2026, pages 28983–28996

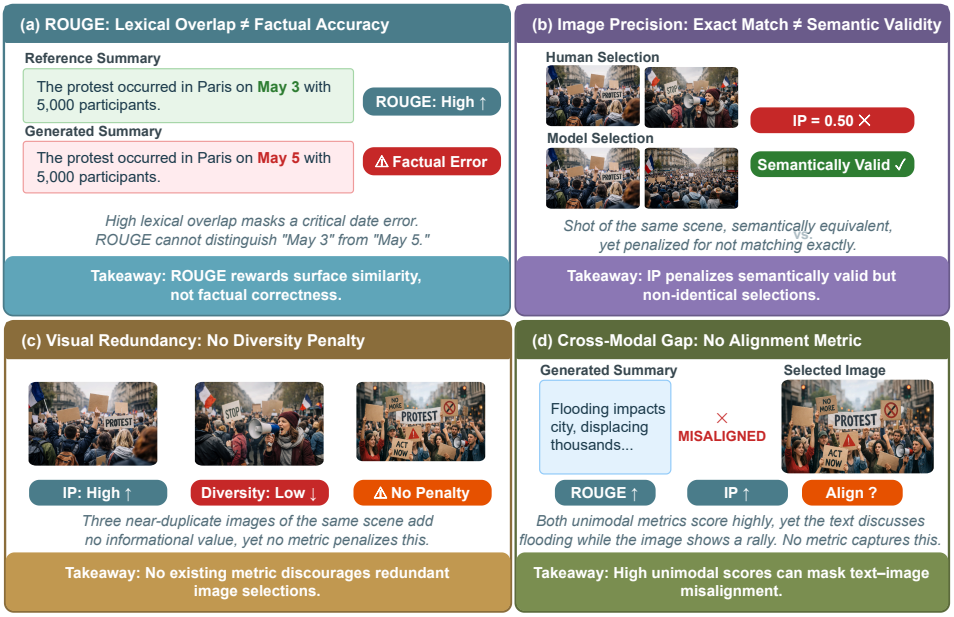

Figure 1: Toy examples showing limitations of current MSMO evaluation.

To address this, we introduce `MM-Eval`, a modular and unified evaluation framework designed to assess multimodal outputs along three key dimensions: quality, alignment, and diversity. Rather than combining existing tools, `MM-Eval` redefines evaluation criteria to reflect the inherently cross-modal nature of MSMO.

The framework consists of three pillars:

- **Text Quality** (S_{text}): We adopt a "decompose-then-verify" approach using OpenFActScore (Lage and Ostermann, 2025) to assess atomic factual consistency, and G-Eval (Liu et al., 2023b), an LLM-based protocol for coherence, fluency, and relevance.
- **Cross-Modal Alignment** ($S_{relevance}$): Using an MLLM-as-a-Judge framework (Zhuang et al., 2024), we measure how well selected images semantically support the generated text, capturing the integrated quality of the multimodal summary.
- **Visual Diversity** ($S_{diversity}$): To discourage visual redundancy, we use Truncated CLIP Entropy (TCE) (Ibarrola and Grace, 2024), which measures the distinctiveness of the image set in latent space.

`MM-Eval` aggregates these components via a learned weighting model trained on the *mLLM-EVAL* news benchmark. Our analysis of human judgments from this dataset reveals a text-dominant preference structure in the news domain. In news, factual consistency serves as a key determinant of perceived quality, while visual relevance and diversity provide complementary signals whose contribution is conditional on adequate textual fidelity. We validate these findings through a human evaluation study that confirms annotators value all three dimensions, supporting MM-Eval's multi-pillar design. Since the component scorers are reference-weak and domain-portable, `MM-Eval` can be applied to new domains by recalibrating only the lightweight aggregation weights with minimal human supervision. This paper details the design of `MM-Eval`, its component metrics, and the empirical patterns uncovered through alignment with human ratings.

2 Literature Review

2.1 Traditional Evaluation Metrics

Text-based Metrics. Summarization evaluation has long relied on ROUGE (Lin, 2004), which computes n-gram overlap with reference summaries. While effective for extractive systems, ROUGE cannot capture semantic equivalence, paraphrasing, or factual errors, and penalizes bridging sentences that refer to accompanying visuals (Schluter, 2017). BERTScore (Zhang et al., 2019) addresses lexical rigidity via contextual embeddings but remains reference-based and fails to distinguish factual from hallucinated content (Zhang et al., 2024; Wan and Bansal, 2022).

Image-based Metrics. Image selection is typically evaluated using Image Precision (IP) and Im-

age Recall (IR), which check whether the model selects the exact annotator-chosen images (Zhu et al., 2020). This assumes a single correct reference set: if a model selects a semantically equivalent but different image, it receives no credit. Moreover, redundant images in the reference set may inflate IP without adding informational value.

Multimodal Aggregation. The Multimodal Automatic Evaluation (MMAE) framework (Zhu et al., 2018) aggregates ROUGE-L, IP, and image–text cosine similarity via a regression model trained on human satisfaction scores. While the aggregation principle is sound, reliance on ROUGE and IP retains the shortcomings mentioned above. `MM-Eval` builds on this idea but incorporates semantically grounded, reference-weak components.

2.2 LLM-based Evaluation Approaches

LLM-as-a-Judge. Recent work such as *mLLM-Eval* (Zhuang et al., 2024) uses large multimodal models (e.g., GPT-4V) to rate summaries directly via Chain-of-Thought prompting. While effective, this approach is resource-intensive and acts as a black box with no explicit quality breakdown. `MM-Eval` addresses these gaps by decomposing evaluation into interpretable, modular sub-metrics built on open-source models, allowing individual components to be independently replaced or upgraded.

Fact-level and Diversity Evaluation. FActScore (Min et al., 2023) proposes a "decompose-then-verify" approach, breaking summaries into atomic facts validated against the source. OpenFActScore (Lage and Ostermann, 2025) extends this with open-source models. For visual diversity, Truncated CLIP Entropy (TCE) (Ibarrola and Grace, 2024) computes entropy over eigenvalues of CLIP embeddings, providing a reference-free signal that penalizes redundant visual outputs without requiring large sample sets like FID (Heusel et al., 2017). The extent to which spectral entropy in CLIP space fully captures human notions of visual diversity remains an open question.

3 Methodology

Formally, the input to `MM-Eval` consists of a source document $D = \{T_{\text{source}}, V_{\text{source}}\}$ and a system-generated summary $S_{\text{cand}} = \{T_{\text{gen}}, V_{\text{sel}}\}$. The output is a scalar score $S_{\text{final}} \in \mathbb{R}$, representing the overall quality of the candidate summary.

The score is computed by integrating signals from three core components: (1) textual quality, (2) cross-modal alignment, and (3) visual diversity.

3.1 Pillar 1: Quantifying Textual Summary Quality (S_{text})

The textual dimension of `MM-Eval` assesses two core aspects: **Factual Consistency**: whether the generated statements are verifiable with the source; and **Qualitative Attributes** such as coherence, relevance, and fluency.

3.1.1 Factual Consistency

To identify hallucinations and ensure faithfulness to the source, we adopt the OpenFActScore pipeline, which shifts evaluation from n-gram recall to fact-level precision.

Step 1: Atomic Fact Generation (AFG). The generated summary T_{gen} is decomposed into a set of atomic facts $A = \{a_1, a_2, \ldots, a_m\}$, where each a_i expresses a single verifiable claim (e.g., "The event occurred on Tuesday") (Lage and Ostermann, 2025). We employ an instruction-tuned open-source LLM for this step.

Step 2: Atomic Fact Validation (AFV). Each atomic fact a_i is then validated against the source text T_{source} using a second LLM, which functions as a binary classifier.

Step 3: Factuality Scoring. The factual consistency score is computed as the average precision over validated facts:

$$S_{fact} = \frac{1}{|A|} \sum_{i=1}^{|A|} v_i \tag{1}$$

where $v_i \in \{0, 1\}$ indicates whether a_i is supported by the source. The resulting score ($0 \leq S_{\text{fact}} \leq 1$) is robust to paraphrasing and insensitive to output length (Min et al., 2023).

3.1.2 Qualitative Metrics

Subjective qualities such as relevance (S_{rel}), coherence (S_{coh}), and fluency (S_{flu}) are difficult to quantify using explicit formulas. We build on G-Eval (Liu et al., 2023b), an LLM-as-a-Judge approach shown to correlate strongly with human judgments. Standard LLM prompting can exhibit high variance (e.g., oscillating between adjacent scores), which G-Eval addresses through a structured form-filling protocol with Chain-of-Thought (CoT) and probability-weighted scoring.

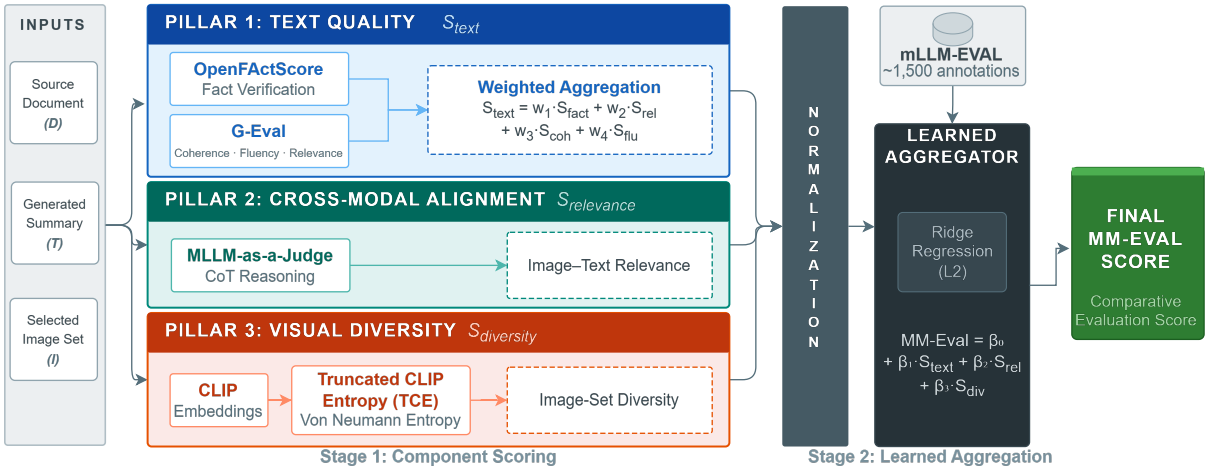

Figure 2: An overview of MM-Eval Framework.

Finally, S_{text} is a weighted aggregation of all four intra-pillar components as below:

$$S_{text} = w_1 S_{fact} + w_2 S_{rel} + w_3 S_{coh} + w_4 S_{flu} \quad (2)$$

3.2 Pillar 2: Assessing Image-to-Text Relevance ($S_{relevance}$)

This pillar measures the alignment of the modalities. It answers the question: Do the selected images actually complement/supplement the text?

3.2.1 MLLM-as-a-Judge for Alignment

`MM-Eval`'s MLLM-as-a-Judge methodology captures the complex pragmatic relevance required by human annotators. Recent advancements show that MLLMs are capable of higher correlation with human relevance judgments than standard proxies. The MLLMs can be instructed to perform Chain-of-Thought (CoT) reasoning before assigning a score, mitigating the high variance observed in human preference (Sahili et al., 2025).

By stabilizing the measurement of cross-modal alignment, this component ensures the metric robustly evaluates whether the images truly add value to the summarized facts. Since `MM-Eval` is modular by design, the specific MLLM used for alignment scoring can be replaced as stronger open-source or proprietary models become available, without altering the rest of the framework.

3.3 Pillar 3: Measuring Visual Diversity ($S_{diversity}$)

A common failure mode in summarization is visual redundancy. For example, a news article about a protest might include multiple images captured from slightly different angles or taken seconds apart, each visually similar but not meaningfully distinct. A model might select several such near-duplicate images, resulting in low informational diversity. `MM-Eval` explicitly penalizes this behavior using spectral entropy encouraging more diverse and informative visual outputs.

3.3.1 Truncated CLIP Entropy (TCE)

Pillar 3 addresses the common multimodal failure mode of visual redundancy, and employs TCE to quantify $S_{diversity}$ using information theory.

The calculation of TCE involves generating CLIP embeddings (F) for the selected images and computing their empirical covariance matrix (C). The eigenvalues (λ_i) of C represent the distribution of variance of the selected images across the principal axes of the semantic feature space (Ibarrola and Grace, 2024). By normalizing these eigenvalues into a probability distribution (p_i) and then calculating the Von Neumann entropy

$$S_{diversity} = -\sum_{i=1}^{k} p_i \log(p_i), \quad (3)$$

the metric effectively measures the semantic volume spanned by the selected images in the CLIP embedding space (Ospanov et al., 2025).

This sophisticated approach is necessary because simple pixel-based or pairwise distance metrics fail when images are semantically redundant but visually distinct (e.g., two images of the same static scene taken seconds apart with minor lighting changes). By operating in the semantic feature space defined by CLIP, TCE ensures that diversity is penalized only when the selected images convey highly overlapping semantic content. We note, however, that spectral entropy in CLIP space serves as a proxy for human notions of visual diversity; the degree to which it captures all perceptually

meaningful distinctions remains subject to further validation.

3.4 Synthesizing the Composite Score

The final step is to combine S_{text}, $S_{relevance}$, and $S_{diversity}$ into a single `MM-Eval` score. A simple average is inappropriate, as these sub-metrics differ in both scale and distribution (S_{text} is a precision probability, S_{rel} a Likert score, and S_{div} a log-entropy value).

3.4.1 The Regression Model

In order to learn the weights for all the `MM-Eval` components, we adopt the supervised learning approach pioneered by MMAE. We treat the aggregation as a regression problem where the goal is to predict the human judgment score.

Training Data: We utilize the *mLLM-Eval* Benchmark dataset. This dataset contains:

- Inputs: 142 distinct multimodal news articles.
- Outputs: Summaries generated by 9 different models (seq2seq, various MLLMs).
- Labels: $\sim 1,500$ expert annotations providing "Overall Quality" scores (y_{human}).

Model: We define the `MM-Eval` score as a linear combination of the normalized sub-scores:

$$\text{MM-Eval} = \beta_0 + \beta_1 S_{text} + \beta_2 S_{relevance} + \beta_3 S_{div} \tag{4}$$

We define the feature vector $X_i =$ for each sample i in the benchmark. We then solve for the weights $\vec{\beta}$ that minimize the Mean Squared Error (MSE) against the human scores y_{human}:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i} \left(\beta^T X_i - y_{human}^{(i)}\right)^2 \tag{5}$$

Since the learned weights $\hat{\beta}$ reflect the preference structure of a specific evaluation context (here, news summarization), they may not directly transfer to domains where visual information plays a more central role. However, the aggregation model is deliberately shallow and low-dimensional, requiring only a small set of human preference judgments (e.g., ordinal ratings on tens to a few hundred examples) to recalibrate for a new domain. Crucially, the underlying component scorers, OpenFActScore, G-Eval, the MLLM judge, and TCE, remain applicable without modification, as they operate on the source document and system output alone and do not depend on domain-specific reference summaries.

4 Analysis and Results

4.1 Implementation Details

To obtain individual components scores for the three pillars of `MM-Eval`, we used Mistral-7B-Instruct for S_{text}, LLaVA-Mistral for $S_{diversity}$, and ViT for $S_{relevance}$. For learning the weights among components, we used Ridge Regression (with L2 regularization) and learned the weights in two stages, first stage for the different components involved in S_{text} and then for the three components involved in Eq. (4) in stage 2. All scoring components used deterministic decoding (temperature 0, sampling disabled) to ensure reproducibility; further technical details are listed in Section A.1.[1]

4.2 Statistical Analysis of Annotations

To rigorously justify the structure and parametrization of `MM-Eval`, it is necessary to analyze the underlying statistical properties of the human evaluation dataset. The descriptive statistics and visual distributions of the human scores reveal key patterns in evaluator consensus, model performance characteristics, and the points of difficulty in judging summary quality.

Metric	Mean (μ)	Std. Dev. (σ)
Fluency	4.156	0.631
Coherence	4.104	0.664
Image-Set Quality	3.871	0.709
Overall Quality	3.672	1.002
Text–Images Relevance	3.689	1.224
Relevance	3.574	1.295
Consistency	4.044	1.412

Table 1: Human evaluation descriptive statistics (N=1562).

The human-annotations conducted by *mLLM-Eval*, based on $N = 1,562$ samples, provide a detailed look into how annotators perceived quality across seven dimensions and the same is described in Table 1.

Linguistic Quality Consensus: Metrics concerning basic linguistic quality, specifically Fluency ($\sigma = 0.631$) and Coherence ($\sigma = 0.664$), as shown in Table 1, exhibit relatively low variance compared to other evaluated dimensions. The high mean scores ($\mu \approx 4.1$) combined with this limited dispersion indicate that system outputs are generally rated favorably along these dimensions, with comparatively little separation between models. As a result, fluency and coherence appear to function

[1]Code: https://github.com/abidmeeraj/MM-Eval

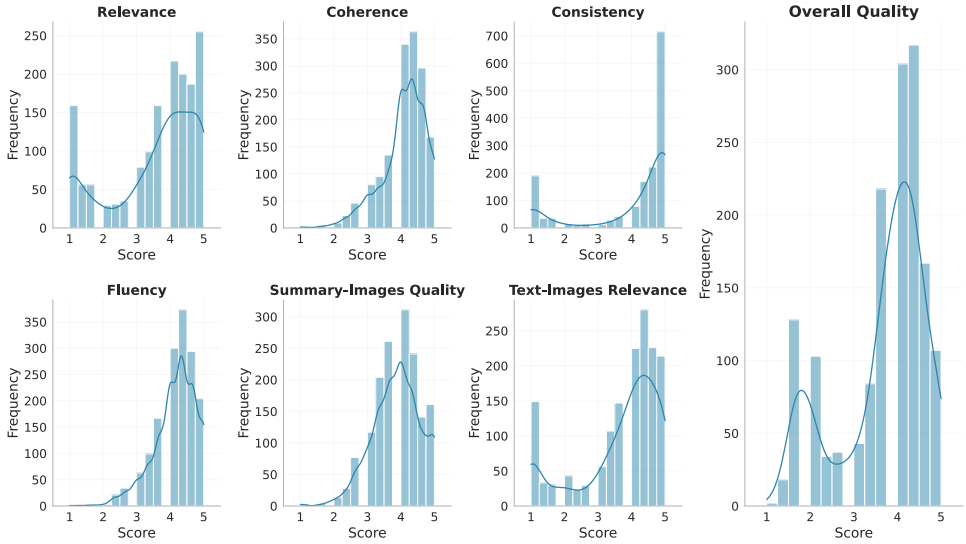

Figure 3: Human evaluation score distributions.

more as baseline quality indicators than as primary differentiators among systems in this evaluation setting.

The Intractability of Consistency and Relevance: By comparison, dimensions related to information alignment and accuracy (Consistency ($\sigma = 1.412$), Relevance ($\sigma = 1.295$), and Cross-Modal Relevance ($\sigma = 1.224$)), as shown in Table 1, exhibit substantially higher variance than linguistic quality metrics. This increased dispersion is consistent with prior findings that factual grounding and content selection are more challenging to assess reliably in summarization tasks (Yuan et al., 2024). In particular, the relatively large standard deviation observed for Consistency indicates greater variability in the assigned scores, which the regression model accounts for by assigning increased importance to S_{fact} during aggregation.

4.3 Interpretation of Score Distributions

The distribution plots, given in Figure 3, offer critical evidence explaining the regression model's learned hierarchy. While Fluency and Coherence exhibit distributions heavily concentrated toward the high end (suggesting ceiling effects), the distributions for Consistency, Relevance, and Overall Quality are widely spread or exhibit complex structures.

Bimodal Distribution of Factual Consistency The distribution plot for Consistency displays a pronounced bi-modal pattern, characterized by significant peaks at both the low end (scores 1-2) and the high end (scores 4-5). This distribution suggests two fundamentally different categories of model output: summaries that are factually sound (high scores) and summaries that contain significant, detectable hallucinations (low scores).

This bi-modal structure empirically supports the hypothesis of hallucination as a deal-breaker behavior in human evaluation. When a factual error is detected, the quality score plunges, regardless of how well the summary performs on secondary criteria like fluency or image quality. Conversely, outputs that maintain high factual integrity receive high scores. This clear separation confirms that Factual Consistency acts as a necessary gate function for overall quality.

This gatekeeping effect is further illustrated in Figure 4, where both human consistency bins (panel a) and automatic factual consistency quintiles (panel b) exhibit a sharp crossover: as consistency increases, the probability of a high overall rating ($\geq$4) rises steeply while the probability of a low rating ($\leq$2) drops to near zero, reinforcing factuality as a critical determinant of perceived overall quality.

The Ridge regression model, by assigning a higher coefficient to S_{fact}, learns to mimic this penalty mechanism, ensuring that a low S_{fact} score, corresponding to the low cluster in the human distribution, translates into a low `MM-Eval` score, mirroring the human response to fabricated information. This non-linear, threshold-like human response motivates a weighted metric that sharply penalizes factual failures rather than smoothing them away via uniform averaging.

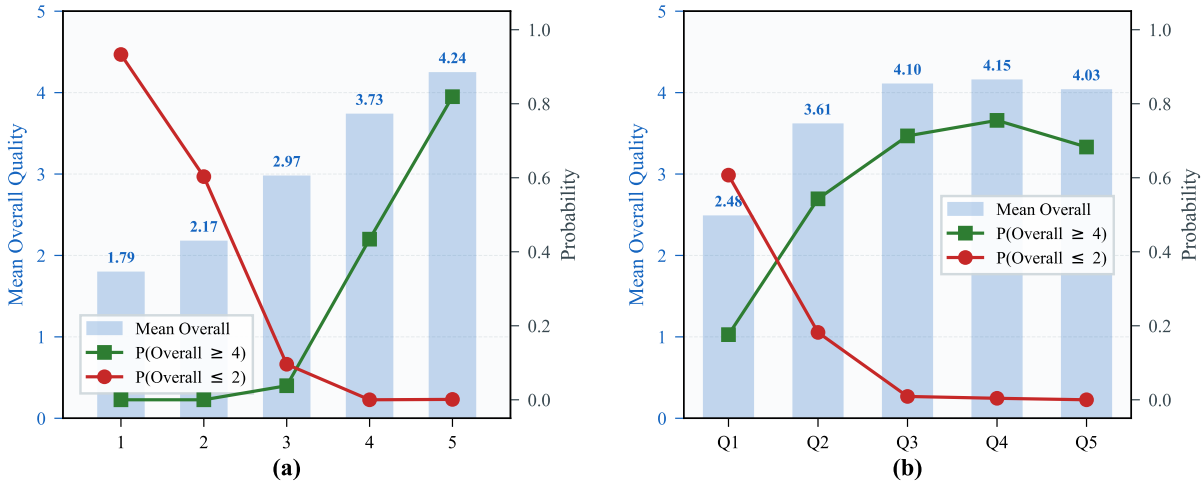

Figure 4: The gatekeeper effect of factual consistency on perceived overall quality. (a) Human consistency bins versus overall quality ratings. (b) Automatic factual consistency quintiles versus overall quality. In both panels, bars indicate mean overall quality (left axis). The green line tracks the probability of receiving a high overall rating ($\geq$4), while the red line tracks the probability of receiving a low overall rating ($\leq$2) (right axis). The crossover pattern illustrates that once factual consistency falls below a threshold, the likelihood of a poor overall score rises sharply regardless of performance on other dimensions.

4.4 The Hierarchy of Weights: Interpretation and Implications

The application of Ridge regression on the human-annotated dataset yields a crucial structural understanding of how annotators prioritize different quality dimensions in multimodal summarization. The resulting weight hierarchy is the most actionable output of the `MM-Eval` analysis, providing direct guidance for system optimization.

Figure 5(a) visualizes the learned aggregation weights alongside per-component rank correlations, contrasting them against an equal-weighting baseline. Equal weighting nearly eliminates rank agreement with human judgments ($\tau = 0.041$), indicating that annotators implicitly apply highly non-uniform priorities across dimensions; this motivates learning the aggregation weights rather than assuming uniform importance.

Table 2 then reports the overall rank correlations and predictive fit statistics of the resulting model. We additionally verified that the learned weight ordering is stable under resampling: across repeated train/test splits, the dominance of S_{text} and the centrality of factual consistency within text remain consistent, indicating the hierarchy is not an artifact of a single split.

We report rank correlations (τ, ρ) because the downstream use of `MM-Eval` is comparative evaluation; absolute calibration is secondary to preserving human ranking preferences. Accordingly, the learned aggregation does not simply improve predictive performance, but more closely reflects the implicit trade-offs that humans apply when combining multiple criteria into a single overall judgment. Given the moderate magnitude of these correlations, `MM-Eval` is best suited for coarse-to-medium system comparisons; when systems are extremely close in quality, targeted human evaluation remains advisable.

Metric	Value	95% CI
Kendall's τ	0.3744	0.374 [0.300, 0.444]
Spearman's ρ	0.5139	0.514 [0.417, 0.597]
Pearson's r	0.6114	–
R^2 (Test Set)	0.3719	–
RMSE (Test Set)	0.8281	–

Table 2: Overall correlation and error metrics with bootstrap confidence intervals.

4.4.1 Learned Weights and The Textual Imperative

The learned aggregation weights shown in Figure 5(a) indicate a strong dominance of the textual component, with Text Quality (S_{text}) accounting for 79% of the total aggregate weight. This is consistent with prior work emphasizing text as the core narrative medium in news summarization (Zhu et al., 2020). Since Ridge Regression shrinks coefficients, lower values do not imply less importance and vice-versa (James et al., 2023). This hierarchy should not be interpreted as evidence that visual

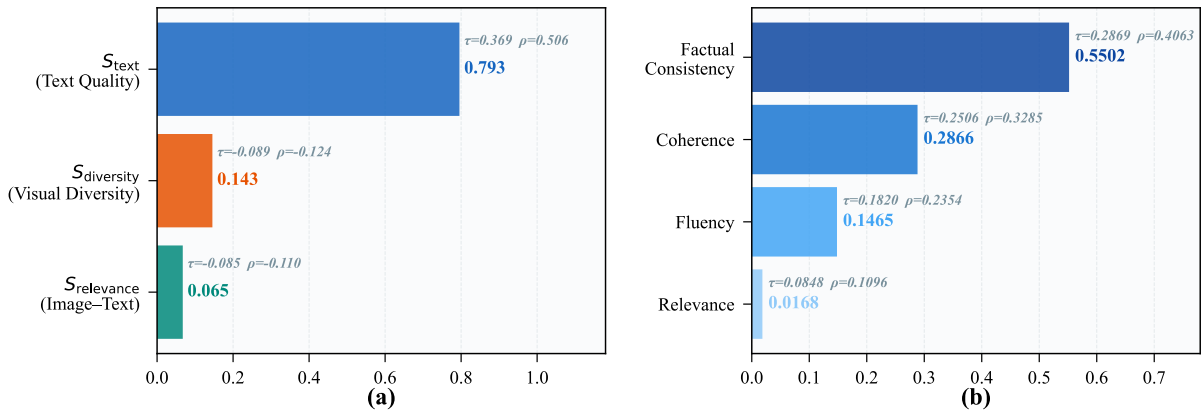

Figure 5: Learned weight hierarchy of `MM-Eval` components. (a) Pillar-level weights with individual rank correlations (τ, ρ) against human judgments. (b) Intra-text component weights within S_{text}.

Component	Relative Contribution (%)
Factual Consistency	$\approx 43.5\%$
Other Text Qualities	$\approx 35.5\%$
Text Quality (S_{text})	79.0%
Image-to-Text Relevance	$\approx 15.0\%$
Visual Diversity (TCE)	$\approx 6.0\%$

Table 3: Learned component weights and their relative contribution to the $MM\text{-}Eval$ score.

Metric	Kendall's τ	Spearman's ρ
Factual Consistency	0.2869	0.4063
Coherence	0.2506	0.3285
Fluency	0.1820	0.2354
Relevance	0.0382	0.0377
Image-Text Relevance	**-0.0848**	**-0.1096**
Image-Set Diversity	**-0.0891**	**-0.1242**

Table 4: Correlation strength of individual metrics.

dimensions are unimportant in general; rather, it reflects the marginal contribution of each component to overall quality judgments in a domain where text carries the primary informational load.

Figure 5(b) further decomposes S_{text} into its constituent dimensions, revealing that factual consistency contributes the largest share of the text-level weight (0.55), followed by coherence (0.29), fluency (0.15), and relevance (0.02).

4.4.2 Implications for Multimodal System Design

The weight hierarchy derived from `MM-Eval` offers a data-driven view of how different quality dimensions contribute to human judgments in multimodal summarization. The learned weights suggest an implicit prioritization: ensuring factual grounding, followed by improvements in textual structure and coverage, and finally refinements to visual alignment and redundancy reduction, whose benefits appear conditional on strong textual quality.

Consistent with this ordering, Table 3 shows that factual grounding accounts for the largest share of the aggregate score. Once textual quality is established, cross-modal alignment ($S_{relevance}$) contributes more substantially than aesthetic refinements such as visual diversity ($S_{diversity}$), which

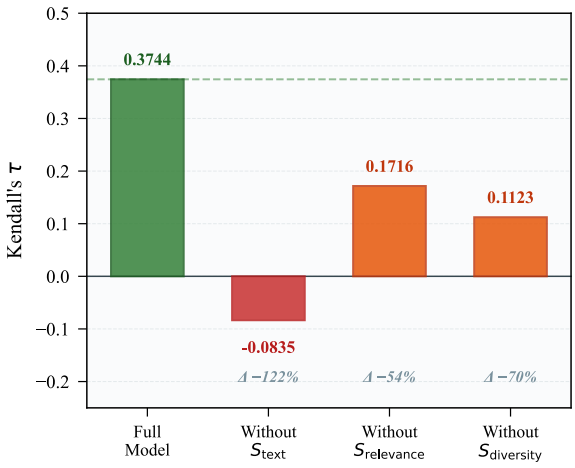

Figure 6: Ablation study showing each pillar's contribution to agreement with human rankings (Kendall's τ). Removing S_{text} causes agreement to drop below zero ($\Delta = -122\%$), while removing visual components also leads to substantial degradation.

exhibit comparatively smaller marginal effects in this domain.

Importantly, these priorities cannot be inferred from pairwise correlations alone. Although Table 4 reports weak or negative associations for the visual proxies, the ablation results shown in Figure 6 indicate that removing $S_{\text{relevance}}$ or $S_{\text{diversity}}$ leads to a marked drop in agreement with human rankings.

This suggests that visual components contribute in a conditional or interaction-dependent manner, rather than through monotonic effects.

One plausible explanation is that, in text-dominant news summarization, visual redundancy or topical similarity does not consistently translate to perceived usefulness once the narrative is already conveyed by the text, and may in some cases align with less informative visual choices. More broadly, the discrepancy between marginal and joint effects is consistent with a conditional contribution pattern: visual components primarily refine quality distinctions when textual fidelity is already adequate, but may co-vary with lower-quality outputs when considered in isolation (e.g., systems that select topically similar but uninformative images may also produce weaker text). This conditional structure explains why removing these components degrades the joint model (Figure 6) despite their weak or negative marginal correlations (Table 4), and underscores the importance of evaluating multimodal quality jointly rather than dimension by dimension.

4.5 Human Evaluation

A potential confound in interpreting the learned weight hierarchy is that the relatively low correlations observed for visual components may reflect limitations of the automatic proxies (TCE, MLLM judge) rather than genuine human indifference to these dimensions. To disentangle proxy noise from preference structure, we conducted a supplementary human evaluation on 200 randomly sampled benchmark articles. Three independent Amazon Mechanical Turk annotators rated each article, on a 5-point Likert scale, along four dimensions: text quality, image relevance, image diversity, and overall quality.

Table 5 shows that annotators rated the visual dimensions favorably, with image relevance and diversity receiving scores comparable to text and overall quality. This indicates that the learned text-dominant hierarchy should not be interpreted as human indifference to visual quality. Rather, it reflects the *marginal contribution* of each component to overall quality in this benchmark, where factual consistency acts as a gatekeeper. The agreement statistics further suggest that, although individual variation is expected, the annotations are sufficiently consistent for aggregate analysis. At the same time, the weaker automatic correlations for $S_{\text{relevance}}$ and $S_{\text{diversity}}$ may partly reflect proxy noise in TCE and the MLLM judge, motivating future work on stronger visual proxies.

Dimension	Mean	$\geq$4	Agreement
Text	3.90 (0.69)	80.1	49.0 / 90.0
Image Relevance	4.04 (0.80)	76.8	44.3 / 84.0
Image Diversity	3.89 (0.83)	73.2	43.0 / 82.2
Overall	4.00 (0.71)	79.2	45.8 / 85.5

Table 5: Human evaluation results. Mean reports mean score with standard deviation in parentheses; $\geq$4 reports the percentage of ratings at least 4; Agreement reports exact agreement / within-1 agreement across annotators.

5 Conclusion

In this study, we proposed `MM-Eval`, a unified evaluation framework for multimodal summarization that integrates textual quality, cross-modal alignment, and visual diversity. Unlike prior approaches that rely on unimodal metrics or opaque model judgments, `MM-Eval` offers an interpretable, reference-weak alternative that aligns with human preferences through supervised aggregation. Its modular design allows individual component scorers to be independently replaced or upgraded as stronger models become available.

Our analysis highlights a text-dominant hierarchy in the evaluation of news summaries, with factual consistency emerging as a critical determinant of overall quality. A human evaluation confirms that annotators value visual relevance and diversity, indicating that the text-dominant weighting reflects the marginal contribution structure of the news domain rather than human indifference to visual dimensions. These findings validate the design of `MM-Eval` and offer practical guidance for system development: prioritizing factual grounding before optimizing visual components is likely to yield the largest gains in perceived quality within text-dominant settings.

Limitations

While `MM-Eval` is validated on a text-dominant news summarization dataset, the learned weight hierarchy is shaped by domain-specific human evaluation behavior. In particular, the strong emphasis on factual consistency reflects the high variability and non-linear judgment patterns observed in this setting, where factual errors are heavily penalized relative to other dimensions. Furthermore, since most of the existing metrics rely on the reference-summaries as important component for evaluation, future work should assess the quality of reference

summaries in existing datasets, in light of the new text-quality dimensions introduced.

The component scorers (OpenFActScore, G-Eval, the MLLM judge, and TCE) are reference-weak and domain-portable, as they operate on the source document and system output alone. When applying MM-Eval to a new domain, the learned aggregation weights from the news setting can serve as a reasonable default (zero-label transfer); if the target domain exhibits a substantially different preference structure, recalibrating only the aggregation model with a small set of human preference judgments (e.g., ordinal ratings on tens to a few hundred examples) provides a lightweight adaptation path. Future work should investigate the behavior of the framework in image-dominant domains (e.g., product reviews or technical documentation), where visual information plays a more central role. Re-calibrating the aggregation weights via Ridge regression in such settings would test whether MM-Eval can dynamically adapt its prioritization of modalities in response to different human evaluation contexts, further assessing its generality.

The automatic proxies used for visual relevance ($S_{\text{relevance}}$) and diversity ($S_{\text{diversity}}$) may not fully capture human perceptions of these dimensions. As shown in our human evaluation (Section 4.5), annotators rate visual quality favorably, suggesting that the weaker automatic correlations for these components partly reflect proxy noise. Replacing TCE or the MLLM judge with stronger visual evaluation models may improve the fidelity of these pillars and shift the learned weight distribution.

Given the moderate magnitude of the observed rank correlations ($\tau = 0.374$), MM-Eval is best suited for coarse-to-medium system comparisons. When systems are extremely close in quality, complementary targeted human evaluation remains advisable.

MM-Eval relies on multiple open-source LLMs and MLLMs, which increases computational cost relative to single-score lexical metrics. However, since evaluation is typically performed offline and far less frequently than model training or inference, this overhead is generally acceptable when reliable multimodal assessment is the goal. The modular design also permits practical trade-offs: users may substitute smaller or faster models for individual components, omit optional pillars (e.g., diversity) when resources are limited, or cache intermediate outputs such as atomic facts and CLIP embeddings to amortize repeated evaluations.

Ethical Considerations

Our evaluation framework is trained on human preference data, which may encode subjective biases, such as an overemphasis on textual content. Additionally, reliance on LLMs introduces potential instability and opacity. We recommend using MM-Eval as a complement to, rather than a replacement for, human evaluation, particularly in high-stakes or sensitive domains.

Acknowledgments

This research was undertaken with the assistance of resources from the National Computational Infrastructure (NCI Australia), an NCRIS enabled capability supported by the Australian Government.

References

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

F Ibarrola and K Grace. 2024. Measuring diversity in co-creative image generation. *arXiv preprint arXiv:2403.13826*.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2023. *An Introduction to Statistical Learning with Python*. Springer, New York.

Chaoya Jiang, Rui Xie, Wei Ye, Jinan Sun, and Shikun Zhang. 2023. Exploiting Pseudo Image Captions for Multimodal Summarization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 161–175. Association for Computational Linguistics (ACL).

LF Lage and S Ostermann. 2025. Openfactscore: Open-source atomic evaluation of factuality in text generation. In *arXiv preprint arXiv:2507.05965*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual Instruction Tuning. *Advances in neural information processing systems*, 36:34892–34916.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-EVAL: NLG Evaluation using GPT-4 with Better Human Alignment. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522. Association for Computational Linguistics (ACL).

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-Tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, page 12100.

A Ospanov, M Jalali, and F Farnia. 2025. Scendi Score: Prompt-Aware Diversity Evaluation via Schur Complement of CLIP Embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16927–16937.

Zahraa Al Sahili, Maryam Fetanat, Maimuna Nowaz, Ioannis Patras, and Matthew Purver. 2025. FairJudge: MLLM Judging for Social Attributes and Prompt Image Alignment. In *arXiv preprint arXiv:2510.22827*.

Natalie Schluter. 2017. The limits of automatic summarisation according to rouge. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2, pages 41–45.

David Wan and Mohit Bansal. 2022. Evaluating and Improving Factuality in Multimodal Abstractive Summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9632–9648.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution.

D Yuan, E Rastogi, F Zhao, S Goyal, G Naik, and SP Rajagopal. 2024. Evaluate Summarization in Fine-Granularity: Auto Evaluation with LLM. *arXiv preprint arXiv:2412.19906*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yue Zhang, Jingxuan Zuo, and Liqiang Jing. 2024. Fine-grained and explainable factuality evaluation for multimodal summarization. *arXiv preprint arXiv:2402.11414*.

Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2018. MSMO: Multimodal Summarization with Multimodal Output. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4154–4164, Brussels, Belgium. Association for Computational Linguistics.

Junnan Zhu, Yu Zhou, Jiajun Zhang, Haoran Li, Chengqing Zong, and Changliang Li. 2020. Multimodal summarization with guidance of multimodal reference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9749–9756.

Haojie Zhuang, Wei Emma Zhang, Leon Xie, Weitong Chen, Jian Yang, and Quan Z Sheng. 2024. Automatic, meta and human evaluation for multimodal summarization with multimodal output. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, volume 1, pages 7768–7790.

A Appendix

A.1 Technical Details

We experimented with different models for calculating the scores for the components involved in both the stages. The final models for each component were selected on the basis of best results on the given subset. The models are listed in Table 6.

Component	Model(s)
Factual Consistency	**Mistral-7B-Instruct-v0.1**
	Mistral-7B-Instruct-v0.3
	Llama-2-7b-chat-hf
	Llama-2-13b-chat-hf
Rel, Coh, Flu	**Mistral-7B-Instruct-v0.3**
	Llama-2-7b-chat-hf
	Llama-2-13b-chat-hf
	Phi-3-mini-4k-instruct
	Nous-Hermes-2-Mixtral-8x7B-DPO
Relevance	**Llava-v1.6-mistral-7b-hf**
	Llava-onevision-qwen2-7b-ov-hf
Diversity	**ViT-B/32**
	ViT-B/16
	ViT-L/14
	ViT-L/14@336px

Table 6: Models used for each evaluation component. Final models used are highlighted in bold.

The obtained scores using these models were then normalized to a common range, the details to which are listed in Table 7. Its important to mention that we used TCE for the calculation of image-set diversity with the maximum eigen size of 20.

Prompts: For S_{text}, we used the same prompts as G-Eval (Liu et al., 2023b) except Factual Consistency, for which we relied on the default prompts in OpenFactScorer implementation. An illustration of prompt used for $S_{relevance}$ is given as Figure 7.

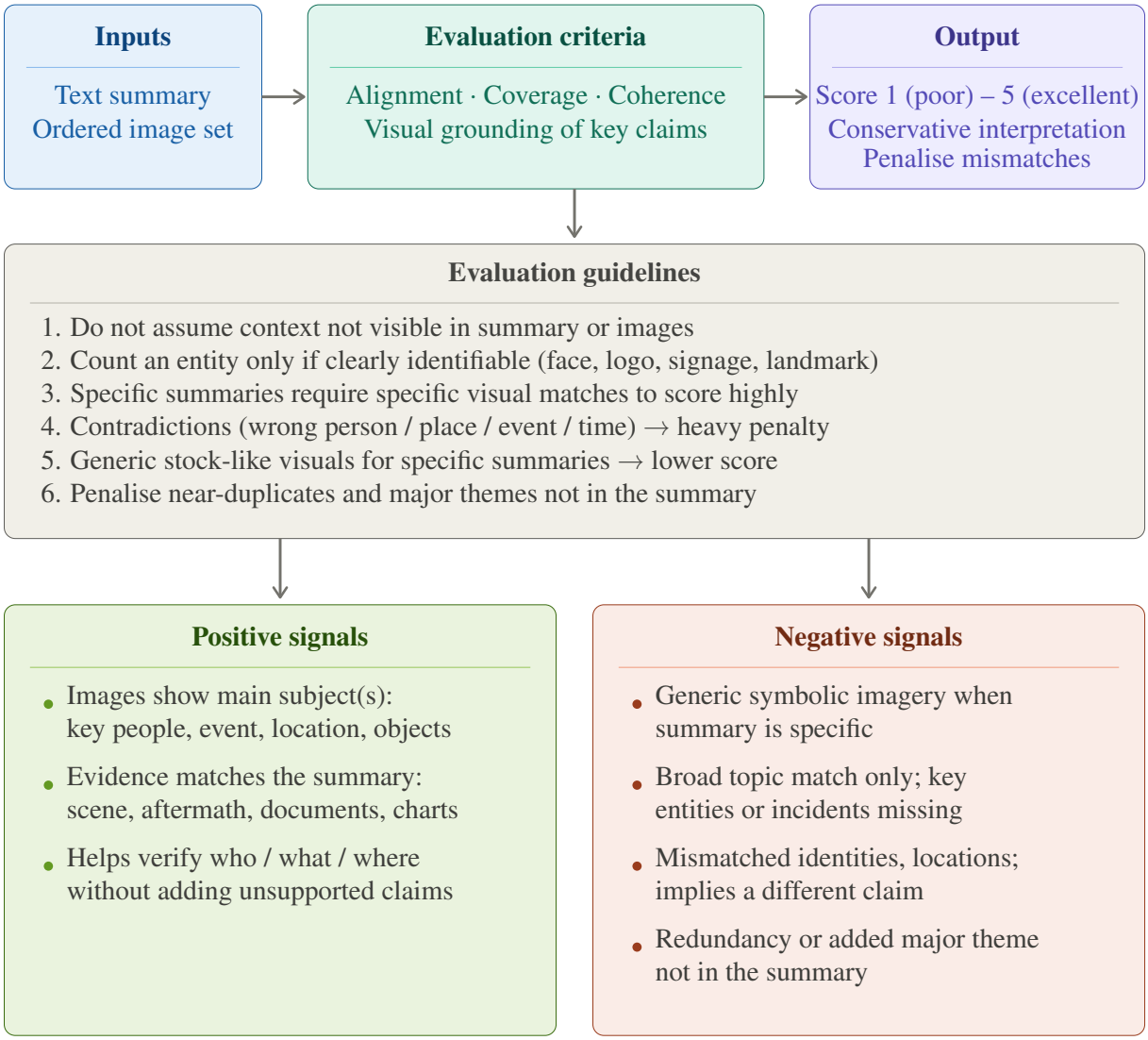

Figure 7: An Illustration of Image-Text Alignment scoring protocol.

Metric	Original	Normalized
Factual Consistency	[0, 100]	[0, 1]
Other Textual Components	[0, 1]	[0, 1]
LLaVA Relevance	[1, 5]	[0, 1]
TCE Diversity	[0, 15]	[0, 1]

Table 7: Normalization bounds applied to individual evaluation metrics.

Metric	Mean	Std	2.5%	97.5%
Kendall's τ	0.3788	0.0251	0.3382	0.4265
Spearman's ρ	0.5176	0.0312	0.4703	0.5732

Table 8: Stability of rank correlation across resampling runs.

Decoding Settings: To ensure reproducibility and minimize run-to-run variance, all LLM- and MLLM-based scoring components (OpenFActScore, G-Eval, and the MLLM-as-a-judge for $S_{\text{relevance}}$) were configured with deterministic decoding: temperature was set to 0 and sampling was disabled. This departs from the original G-Eval protocol, which uses probability-weighted scoring over sampled tokens; we opted for deterministic outputs to prioritize consistency in an offline benchmarking context. The models listed in Table 6 were selected on the basis of best agreement with human judgments on the given subset; since `MM-Eval` is modular, any component model can be substituted without altering the remainder of the pipeline.

As mentioned in Section 4.1, we learned the weights in two stages. For Stage 1 the $\alpha = 1.0$, and for Stage 2 the $\alpha = 0.1$ were selected via a 5-fold cross-validation. The dataset split was 80/20 which was stratified by the summarization system for the summaries in original dataset.[2]

A.2 Stability Analysis

This section provides detailed results about repeated-split end-to-end stability (S=50) where each repetition re-splits, re-learn Stage 1 + Stage 2, and evaluates on its own held-out split. Tables 8, 9, and 10, provide further evidence to the claims that the text-dominant hierarchy is not a single-split artifact: w_{text} remains the largest weight across resamples and factual consistency remains the dominant text component.

A.3 Detailed Numerical Results

Tables 11–15 provide the exact numerical values corresponding to Figures 4–6, in the main text.

[2]The singed coefficients for S_{text}, $S_{relevance}$, and $S_{diversity}$ were 2.7721, 0.2256, and -0.4991 respectively.

Weight	Mean	Std	2.5%	97.5%
w_{text}	0.7572	0.0428	0.6748	0.8306
$w_{\text{relevance}}$	0.0701	0.0215	0.0256	0.1091
$w_{\text{diversity}}$	0.1727	0.0224	0.1431	0.2149

Table 9: Stability of learned pillar weights under resampling.

Text Weight	Mean	Std	2.5%	97.5%
Factual Consistency	0.5511	0.0085	0.5347	0.5644
Coherence	0.2870	0.0109	0.2693	0.3052
Fluency	0.1453	0.0125	0.1242	0.1664
Relevance	0.0166	0.0064	0.0025	0.0262

Table 10: Stability of intra-text weighting coefficients across resampling.

Consistency bin	n	Mean Overall	Median Overall	$\mathbf{P}(\text{Overall} \geq \mathbf{4})$	$\mathbf{P}(\text{Overall} \leq \mathbf{2})$
1	225	1.79	1.67	0.000	0.933
2	58	2.17	2.00	0.000	0.603
3	52	2.97	3.00	0.038	0.096
4	290	3.73	3.67	0.434	0.000
5	937	4.24	4.33	0.819	0.001

Table 11: Relationship between human consistency bins and overall quality ratings.

Factual Consistency quintile	n	Mean Overall	Median Overall	$\mathbf{P}(\text{Overall} \geq \mathbf{4})$	$\mathbf{P}(\text{Overall} \leq \mathbf{2})$
Q1	313	2.48	2.00	0.176	0.607
Q2	313	3.61	4.00	0.543	0.182
Q3	345	4.10	4.00	0.713	0.009
Q4	282	4.15	4.33	0.755	0.004
Q5	309	4.03	4.00	0.683	0.000

Table 12: Relationship between Factual Consistency quintiles and overall quality ratings.

Component	Weight	τ	ρ
Final MM-Eval	–	0.374^*	0.514^*
S_{text} (Text Quality)	0.793	0.369	0.506
$S_{\text{relevance}}$ (Image–Text)	0.065	-0.085	-0.110
$S_{\text{diversity}}$ (Visual)	0.143	-0.089	-0.124
Equal Weights Baseline	0.333	0.041	0.058

Table 13: Correlation of model components with human judgments.

Metric	Weight	τ	ρ
Factual Consistency	0.5502	0.2869	0.4063
Coherence	0.2866	0.2506	0.3285
Fluency	0.1465	0.1820	0.2354
Relevance	0.0168	0.0848	0.1096

Table 14: Correlation of each text-quality component with human scores.

Ablation	Kendall's τ	Δ from Full
Full Model	0.3744	–
Without S_{text}	-0.0835	**-122%**
Without $S_{\text{relevance}}$	0.1716	-54%
Without $S_{\text{diversity}}$	0.1123	-70%

Table 15: Ablation study showing each pillar's contribution.