Seeing Through the MIRAGE: Evaluating Multimodal Retrieval Augmented Generation

Anonymous ACL submission

Abstract

We introduce MIRAGE, an evaluation framework for retrieval-augmented generation (RAG) from multimodal sources. As audiovisual media becomes a prevalent source of information online, it is essential for RAG systems to integrate information from these sources into generation. However, existing evaluations for RAG are text-centric, limiting their applicability to multimodal, reasoning intensive settings because they don't verify information against sources. MIRAGE is a claim-centric approach to multimodal RAG evaluation, consisting of INFOF1, evaluating factuality and information coverage, and CITEF1, measuring citation support and completeness. We show that MIRAGE, when applied by humans, strongly aligns with extrinsic quality judgments. We additionally introduce automatic variants of MIRAGE and three prominent TextRAG metrics–ALCE, ARGUE, and RAGAS–demonstrating the limitations of text-centric work and laying the groundwork for automatic evaluation. We release open-source implementations[1] and outline how to assess mulitmodal RAG.

1 Introduction

People increasingly consume information online through audiovisual media, from firsthand video footage of natural disasters to professional news coverage of major political events. It is thus essential that systems for retrieval augmented generation be able to integrate information from any modality and provide accurate citations to those sources. However, works that focus on generating long form text from multimodal sources (Krishna et al., 2017; Liu et al., 2023a; Lin et al., 2024; Martin et al., 2025) rely on text-only metrics for evaluation (e.g., Lin, 2004; Zhang et al., 2019), that don't verify information against the information sources.

We introduce MIRAGE, a framework for **M**ultimodal **R**etrieval **A**ugmented **G**eneration

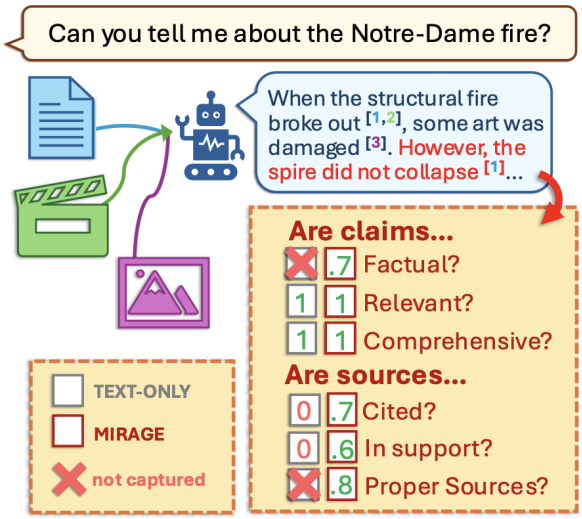

Figure 1: MIRAGE asseses predictions at the claim level, evaluating information factuality and coverage, and citation support and attribution, enabling RAG evaluation grounded in text, image, audio, and video.

Evaluation. Inspired by factuality evaluations (Min et al., 2023), the core of MIRAGE is the idea that all information—whether represented in text, image, audio, or video—can be decomposed into unimodal linguistic objects (subclaims), thus grounding all modalities to subclaims. Our metric consists of INFOF1, to capture the factuality (precision) and coverage of information (recall) in generated text, and CITEF1, to capture if sources cited by the generated text are fully supported (precision) and all the evidence is fully cited (recall). We design this metric to be evaluated in reference-free (no gold human text) and reference-based (with gold human text) settings, allowing for our method to apply to summarization, VQA, and settings without human annotated gold data.[2]

In addition to MIRAGE, we transition three prominent text-based RAG (TextRAG) metrics—ALCE (Gao et al., 2023), ARGUE (Mayfield et al., 2024), and RAGAS (Es et al., 2024)—to multi-

[1]Data in supplementary material, code upon publication.

[2]We elaborate on evaluating other tasks in Appendix I.

modal RAG. However, these methods were not designed for evaluating across modalities or with long multimodal contexts. We find that these existing methods are limited by computational constraints of multimodal inference (e.g. the concatenation of multiple videos in ALCE), have infeasible mechanisms in a visual domain (e.g. extracting all the information from a source in RAGAS), and largely ignore the literature in factuality assessment, like FActScore (Min et al., 2023) (all metrics).

We focus our analysis on evaluating video-based RAG (VideoRAG) in WikiVideo (Martin et al., 2025), the task of writing a Wikipedia-style article in response to a query from multiple videos with citations to the videos used in generation. Video poses various challenges for RAG evaluation, notably around assessments of support (Liu et al., 2020; Zaranis et al., 2025). Judging entailment between a text premise and audio-visual evidence is especially difficult, requiring strong cross-modality understanding. Additionally, performing inference over multiple videos or long videos is often computationally infeasible (Li et al., 2025). These constraints make video a challenging source domain—more so than images or audio taken independently—and therefore the ideal testbed for modality-agnostic evaluation.

In our analysis of the four metrics, we collect three human judgments: (1) extrinsic quality judgments, where human annotators assess the overall quality of predictions; (2) intrinsic quality judgments, where annotators provide INFOF1-specific judgments; and (3) grounding judgments, where annotators ground subclaims from predictions in the relevant videos. Comparing metric agreement with these judgments, we find (1) MIRAGE annotated by humans has the highest agreement with extrinsic quality judgments, (2) TextRAG metrics poorly or negatively correlate with all human judgments (due to the aformentioned limitations) and (3) ROUGE and BERTScore perform well when judging the information presented in predictions, but less so for judgments about groundedness. Based on these findings, we provide recommendations for how to comprehensively evaluate multimodal RAG in time- and compute-constrained settings.

Our contributions can be summarized as follows: (1) We propose MIRAGE for evaluating multimodal RAG. (2) We transition three prominent TextRAG metrics to multimodality. (3) We perform analysis on the agreement between the metrics and human judgments

2 Related Work

There are several works that tackle the task of video-based RAG (VideoRAG; Jeong et al., 2025; Martin et al., 2025; Ren et al., 2025). However, we focus our evaluation on WikiVideo (Martin et al., 2025) because, unlike other work, the task requires generating long-form text, from a well defined information need, and across multiple videos, which makes properly citing sources a task requirement. This best mirrors the common setting of TextRAG with targeted information seekign queries, multiple information sources, and citations to the information used in generation (Barham et al., 2023; Gupta et al., 2024; Han et al., 2024; Lawrie et al., 2024).

Claim decomposition methods are a core component of factuality assessments (e.g. Min et al., 2023; Song et al., 2024). Subclaims, that are decomposed from sentences, are argued the appropriate unit of assessment for evidential support because subclaims are easier and less ambiguous to evaluate than sentences. This allows for a finer granularity in the evaluation of individual, automatic facts, instead of evaluating multiple propositions in a single sentence. There are multiple viewpoints on the level of granularity of subclaims (e.g. Gunjal and Durrett, 2024; Wanner et al., 2024a,b), but we follow the level of subclaim granularity in Martin et al. (2025) in order to utilize their subclaim grounding annotations to evaluate the performance of LLMs and VLMs in claim verification.

We provide individual sections and appendices to the RAG metrics we implement, but other metrics for evaluating the quality of generations exist in the literature. For example, Jing et al. (2024) evaluates the factuality of claims in a caption based on their corresponding image, Nenkova and Passonneau (2004) propose abstracting model summaries into Summary Content Units (SCUs), and Thakur et al. (2025) evaluate levels of support between a sentence and a citation.

3 TextRAG Metrics

Before introducing MIRAGE, we first revisit how RAG is assessed in text-only settings. We look at three prominent metrics—ALCE (Gao et al., 2023), ARGUE (Mayfield et al., 2024), and RAGAS (Es et al., 2024)—providing a foundation for RAG evaluation as well as the modifications to these metrics for multimodality. Implementation details for each metric can be found in Appendix D, Appendix E, and Appendix F, respectively.

Preliminaries We define a *topic* as a subject matter of a *source* (video) or collection of sources. A single source or multiple sources used in generation are the *context*. A sentence is decomposed into a set of *subclaims*–a declarative statement, pertaining to some set of evidence, that may or may not be true. A *prediction* is text produced by a generation system. Anything derived from the prediction we call "predicted X" (predicted {sentence, subclaim, etc.}). A *reference* is some preexisting knowledge on the topic, this can be a knowledge base (Wikipedia Articles; Min et al., 2023), a collection of information relevant to the query (nuggets; Voorhees, 2004), or written text answering the query (reference articles; Martin et al., 2025). We use "reference X" (reference {sentence, subclaim, etc.}) for anything derived from the reference. Finally, a scoring function, e.g. a bidirectional entailment or LLM judgment, on (premise, hypothesis) pairs is used to measure the support between the pair. We denote this as $\mathsf{s}(p, h)$.

3.1 ALCE

ALCE (Gao et al., 2023) aims to capture three qualities of predictions: information factuality and coverage, citation quality, and fluency. To evaluate these qualities, ALCE introduces three metrics: Correctness, Citation Quality, and Fluency.[3]

Correctness To evaluate Correctness, ALCE introduces recall and precision metrics for information. For ELI5 (Fan et al., 2019), a long-form QA task, recall is introduced as an NLI-based judgment between reference subclaims and the prediction. Precision is introduced as the exact match between a predicted answer and a reference answer. To transition Correctness to multimodality, we do not need to modify ALCE Claim Recall, as the verification of reference subclaims in a prediction is independent of the source modality. We do not implement precision because exact match is unsuitable for long-form generation.

Citation Quality To evaluate Citation Quality, ALCE introduces a recall and precision. Citation Quality Recall measures if the concatenation of all cited sources fully supports the citing sentence. Citation Quality Precision detects irrelevant citations by evaluating if a sentence is still supported by its associated citations after removing one of the sources. Both precision and recall involve comparing a sentence against the concatenation of multiple passages. To transition Citation Quality to multimodality, we modify the support from (sentence,passages) pairs to (sentence,videos) pairs.

Citation Quality has two main limitations. (1) From a factuality perspective, the sentences that ALCE compares to the sources express multiple subclaims, making it harder to evaluate the support between the (sentence,sources) pair and requiring that all subclaims in a sentence be supported. (2) From a multimodal perspective, concatenating multiple sources (e.g., multiple videos) raises issues with computational constraints and the lack of multi-video training data for VLMs.

3.2 ARGUE

ARGUE (Mayfield et al., 2024) look to make judgments about predictions for information and citation support. To evaluate these qualities, ARGUE uses nuggets, a collection of question-answer pairs on the topic, to asses the coverage of information (Nugget Coverage) and calculates the support between sentences and citations (Sentence Support).

Nugget Coverage is the proportion of nugget questions correctly answered by the report, aggregated over sentences. This is similar to ALCE's Claim Recall, but focused on recall of QA pairs rather than claims. To transition this to our task, we generate questions from the subclaims with an LLM to create nuggets. The underlying implementation of Nugget Coverage does not need any modification, as the checking if a nugget is recalled by a prediction is independent of the source modality.

Sentence Support is the proportion of predicted sentences attested by *each* of their citations. AUTO-ARGUE (Walden et al., 2025b) implements this for a single citation in a sentence at a time, which we follow. However, Sentence Support faces the same issue as ALCE from a factuality perspective, needing to verify multi-premise sentences against the information sources.

3.3 RAGAS

RAGAS (Es et al., 2024) looks to evaluate three qualities of predictions: faithfulness to the context, the relevance of the prediction to the information need, and the relevance of the context to the information need. To assess these qualities, Es et al. (2024) introduce Faithfulness, Answer Relevance,

[3]We do not explore fluency in this work, as it is not a common failure mode of modern LMs.

and Context Relevance. Note, context in RAGAS is the sources used to generate the prediction.

Faithfulness To evaluate faithfulness to the context, RAGAS introduces Faithfulness. In Faithfulness, statements, decomposed from longer sentences, are compared against the context for a binary judgment of support. To transition Faithfulness to multimodality, we first interpret statements and their decomposition as subclaims decomposed in claim decomposition. Then, we compare the subclaims against the video content.

Like ALCE, Faithfulness encounters issues in the evaluation of support *if* multiple sources are used in generation. However, unlike ALCE, Faithfulness is not explicitly defined to evaluate the concatenation of sources. Thus, we modify the support judgment to be between a single subclaim and single video from the context.

Answer Relevance To evaluate Answer Relevance, RAGAS first takes the prediction and uses an LLM to generate a possible query that was asked to generate the response. Then, this predicted query is compared against the actual query using an embedding model. Answer Relevance requires no modification to transition to multimodality, as the generation of a query and comparison between two queries is independent of the modality.

Context Relevance To evaluate Context Relevance, RAGAS computes a score for each source in the context as $\frac{\text{extracted sentences}}{\text{total sentences}}$, where the numerator is the number of extracted sentences releated to the query from the source. However, transitioning this metric to multimodality is challenging. Unlike text, multimodal content does not have explicit sentences and images and videos are abundant in information. To transition Context Relevance to multimodality, we extract information from the video related to the query and elicit a detailed summary from a VLM to simulate the denominator.

The three TextRAG metrics—ALCE, ARGUE, and RAGAS—each highlight important aspects of RAG evaluation but share core limitations that restrict their applicability to multimodal domains. All three assess information primarily at the sentence level, making fact verification less precise. They also face scalability issues when applied to multimodal data: ALCE relies on the concatenating multiple sources, which is infeasible with multiple videos; and RAGAS assumes the ability to exhaustively extract information from sources, an unrealistic expectation in the visual domain.

MIRAGE is designed to address these issues, while capturing the core features each metric aims to evaluate. Instead of evaluating at the sentence level, it decomposes predicted sentences into subclaims, enabling more granular and interpretable assessments of factuality and citation quality. Its two components—INFOF1 and CITEF1—capture the quality of information and citations, and can be computed in reference-based and reference-free settings. By supporting both reference-based and single/multi-source inference MIRAGE provides a scalable, claim-centric framework, unifying factuality and citation evaluation across any modality.

4 MIRAGE

MIRAGE (Figure 2) is a claim-centric evaluation framework aiming to capture two main aspects of prediction quality: the **information** is factual and covers the information requested by the query, and **citations** fully support their associated subclaims and are properly attributed to the subclaims. To evaluate information, we introduce INFOF1, which measures the proportion of factual predicted subclaims (precision) and the coverage of reference subclaims (recall). To evaluate citations, we introduce CITEF1, which measures the proportion of subclaims supported by their associated citations (precision) and the proportion of recalled information that cites the proper source (recall). For each metric, we implement a reference-based, evaluating against the reference, and reference-free, evaluating against the sources, variation. In this following sections, we introduce MIRAGE as it should be evaluated without considering the limitations of multi-source inference for reference-free evaluations. We provide reformulations for model limitations (single-source inference), weighted variations, and granular alignment in Appendix B.

4.1 INFOF1: Evaluating Information

We introduce INFOF1 to be evaluated as INFOP, to capture factuality, and INFOR, to measure the coverage of information. Whereas traditional summarization and RAG metrics primarily capture surface-level similarity against a reference or only information coverage, INFOF1 provides an interpretable evaluation of the evaluation (unlike ROUGE and BERTScore) for how faithfully a prediction reflects the underlying evidence (unlike RAG metrics) and information need. By evaluating INFOP—

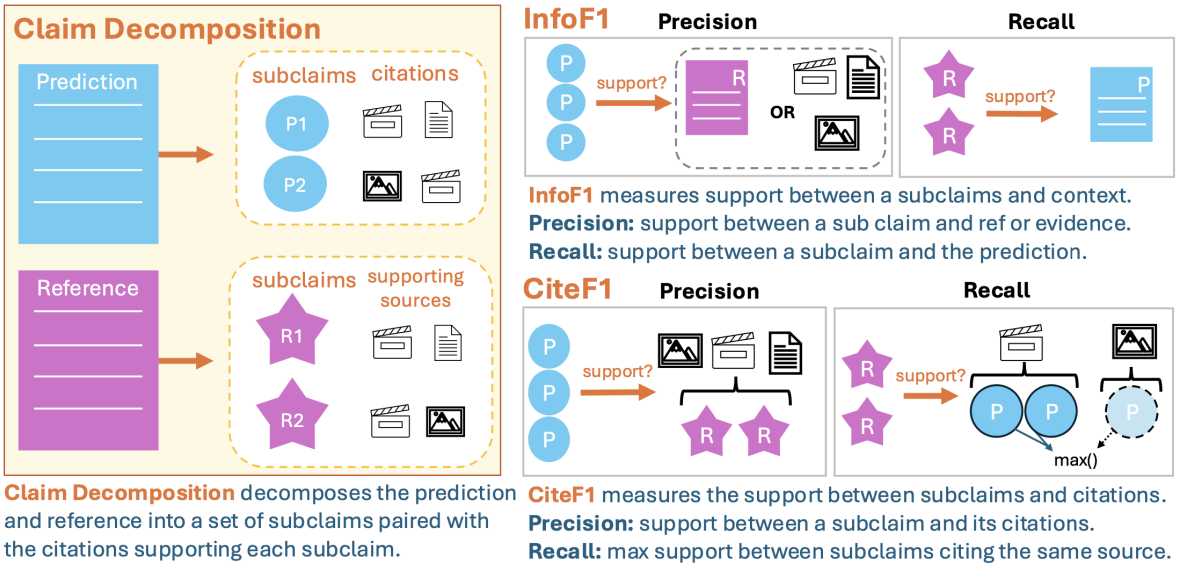

Figure 2: MIRAGE decomposes generated text into subclaims to assess two dimensions of quality: INFOF1, which measures factuality and information coverage, and CITEF1, which measures citation precision and completeness. The framework enables consistent, claim-level evaluation across text, image, audio, and video modalities.

the proportion of predicted subclaims supported by the evidence—and INFOR—the proportion of reference subclaims covered by the prediction—INFOF1 captures if a prediction is factual and fulfills the information need, aligning well with human intuitions and quality judgments (§5).

4.1.1 Information Factuality (INFOP)

To capture the factuality of a prediction (P), we evaluate the factuality of predicted subclaims against a collection of sources ($E = [d_1, \ldots, d_n]$, d_i denoting a source). We first decompose each predicted sentence into a set of subclaims (C_P). These subclaims are then scored for support against the evidence using the scoring function s, evaluating the proportion of supported subclaims:

$$\text{precision}(C_P, E) = \frac{1}{|C_P|} \sum_{c \in C_P} \mathsf{s}(c, E) \qquad (1)$$

We propose two variations of INFOP, Collection Precision and Reference Precision. Collection evaluates subclaims against E, while Reference provides an efficient alternative to E.

Collection Precision is the evaluation of support between subclaims and the collection of evidence, where E is only the videos related to the query topic. It is similar to RAGAS Faithfulness, which evaluates against the sources used during generation (relevant or not). Collection Precision can also be scaled to E as a large knowledge base, becoming a multimodal FActScore (Min et al., 2023).

Reference Precision is the measurement of the predicted subclaims against the reference. We use a human written reference as a proxy for the information in the collection because we consider the human written reference as "extracting" all the relevant information to a query from E. This method provides an efficient alternative to evaluating against E, only requiring evaluation against a single document. However, human annotated references are inherently non-exhaustive for all the possible information that could be included in a prediction. This could lead to penalizing claims as non-factual, when instead they were deemed irrelevant by a human annotator. Thus, reference precision can be considered both evaluating factuality *and* the relevance of the information presented. We offer more on these tradeoffs in Appendix C.

4.1.2 Information Coverage (INFOR)

To ensure that a prediction covers the relevant information from the evidence, we look to evaluate the coverage of reference subclaims against the prediction. We first decompose each reference sentence into a set of subclaims (C_R). Then, we determine a support score for each subclaim relative to the prediction (P):

$$\text{recall}(C_R, P) = \frac{1}{|C_R|} \sum_{c \in C_R} \mathsf{s}(c, P) \qquad (2)$$

Our INFOF1 Recall is the same as ALCE Evaluation of Correctness for ELI5, with both checking if claims in the reference are in the prediction.

Collection Recall Unlike INFOP, we only use the human written reference for recall. References are considered the extraction of all the information relevant to the query from the collection. Collection Recall would require providing an explicit score by inferencing over all the videos and the prediction.

4.2 CITEF1: Evaluating Citations

CITEF1 measures how effectievly a prediction attributes and groudns information to its supporting sources. In any RAG setting, accurate citations are important for transparency–each subclaim must be verifiable. Prior citation metrics overlook this by assessing full sentences against the sources, which obscures notions of support at the subclaim level. We introduce CITEF1 to be evaluated as CITEP—the proportion of subclaims supported by their cited sources—and CITER—the proportion of subclaims that properly attribute the their sources.

4.2.1 Citation Support (CITEP)

To measure if the predicted citations (D_P) support the prediction, we look to evaluate the support between predicted subclaims and their associated citations. We first decompose subclaims and create a mapping between the subclaims and the sources cited by the sentences containing the claims ($c_i : \{d_1, \ldots, d_n\}$. Precision can be expressed as the proportion of subclaims supported by their cited sources ($D_P = D$ for space):

$$\text{precision}(P, D) = \frac{1}{|C_P|} \sum_{c_i \in C_P} \mathsf{s}(c_i, D(c_i)) \quad (3)$$

Like INFOP, we introduce two variations of CITEP, Collection–against the sources in the evidence collection—and Reference—an efficiecnt alterantive to Collection against a proxy text.

Collection Precision is the evaluation of support between subclaims and the cited sources from the evidence collection. This precision variation is similar to ARGUE Sentence Support, however, instead of evaluating the support between a sentence and citation, we evaluate the support between a subclaim and citation.

Reference Precision is the evaluation of support between subclaims and proxy text for the citation's content. To create the proxy, we create a second mapping between the relevant videos for the topic and reference subclaims from the videos ($v_i : \{c_1^R, \ldots, c_m^R\}$). The reference subclaims come from the WikiVideo grounding judgments, which are annotations for the subclaims in the reference and the videos that support them. Thus when evaluating the support Equation 3, we modify the scoring function for the set of reference subclaims.

$$\mathsf{s}(c_i^P, [c_1^R, \ldots, c_m^R]) \quad (4)$$

Prior metrics, ARGUE and ALCE, evaluate full sentences against their citations, but this requires the sentence to be fully supported by each video, either leading to unnatural predictions, with short, single-claim sentences scoring best, or systems being unnecessarily penalized when some of the subclaims would have been supported.

4.2.2 Proper Attribution (CITER)

Here, we aim to determine the proportion of *reference* subclaims that (1) are attested by the prediction via (2) sentences that cite correct sources—i.e. those that *also* attest those subclaims. To do so, we first decompose the reference into subclaims and create a mapping between each subclaim and *all* videos in the collection that attest that subclaim ($c_i^R : \{v_1, \ldots, v_n\}$). Next, we create another mapping from each video cited *in the prediction* to *all* predicted sentences citing that video ($v_i : \{s_1^P, \ldots, s_m^P\}$). We then use these two mappings to score the strength of support for c_i^R provided by $\{s_1^P, \ldots, s_m^P\}$ (denoted as vs_i^P for space), taking the max score over all v_i:

$$\text{recall} = \frac{1}{|C_R|} \sum_{c_i \in C_R} \max_{cv_i \in CV} \mathsf{s}(c_i^R, vs_i^P) \quad (5)$$

We take a max because we care about the *best* support the prediction provides for each c_i^R.

5 Alignment With Human Judgments

Evaluation Setup We generate predictions in three settings for evaluation across 10 WikiVideo topics.[4] (1) **LLM only**. We generate a prediction with an LLM (Qwen3; Yang et al., 2025). Then, every sentence is used to retrieve one video from the oracle video set using OmniEmbed (Ma et al., 2025). (2) **CAG-2 Oracle**. We generate predictions with the CAG pipeline (Martin et al., 2025) using the oracle video set. (3) **CAG-2 RAG**. We generate predictions with the CAG pipeline retrieving from MultiVENT2.0 (Kriz et al., 2025) with MMMORRF (Samuel et al., 2025)

[4]We refer to these as the evaluation topics. 6 covering events from pre-2024 and 4 from events in 2025.

Eval	Ano	INFOP		INFOR	ALCE	ARGUE	RAGAS		R-L	BS
		Ref.	Col.	Ref.	Clm. Rec.	Nugg. Cov.	Faith	Ans. Rel.		
EQJ	1	26.3	23.3	16.3	-5.0	-3.5	32.8	4.8	55.8	52.3
	2	18.2	31.0	31.2	21.2	33.0	29.8	23.3	71.2	81.2
	3	17.8	28.2	13.0	5.0	11.3	16.2	-1.5	41.2	74.2
ICJ	1	38.2	11.5	3.5	27.8	2.2	11.8	44.8	66.7	70.0
	2	26.7	16.3	16.7	23.0	15.2	23.3	33.3	70.0	86.7
	3	26.7	14.8	12.2	11.5	25.2	26.7	23.3	68.2	78.2
GJ	-	6.6	21.5	1.8	-8.2	-4.8	24.8	-26.7	-11.5	0.0

Table 1: Kendall's Tau with all human judgments against the metrics that evaluate information. R-L: ROUGE-L, BS: BERTScore, EQJ: Extrinsic Quality Judgment, ICJ: Intrinsic Claim Judgment, GJ: Grounding Judgment

	EQJ 1	EQJ 2	EQJ 3
ICJ 1	49.0	61.3	54.5
ICJ 2	60.5	81.2	72.7
ICJ 3	65.4	76.3	70.8
GJ	18.2	-5.5	23.0

Table 2: Kendall's Tau agreement between MIRAGE human annotated judgments (ICJ[1-3]) and human annotated quality judgments (EQJ[1-3]).

For the CAG methods, Martin et al. (2025) claim CAG provide citations in generation. However, we find the number of citations limited. Thus, we do post-hoc citations for any sentence without a citation using Video-ColBERT (Reddy et al., 2025).

The underlying scoring function used for evaluation is either an LLM (Qwen 2.5, Qwen et al., 2025) or VLM (Qwen2.5 VL, Bai et al., 2025), in either few-shot (LLM) or zero-shot (VLM).

Collection of Human Judgments To evaluate the various metrics introduced above, we collect three human judgments.[5] (1) **Extrinsic Quality Judgments**. Human annotators are given information about the topic and are asked to give a 1-5 likert considering the following attributes (in order of importance): factuality, adequacy, coherence, relevancy, and fluency. (2) **Intrinsic Claim Judgments**. Human annotators are given the prediction, predicted subclaims, reference, and reference subclaims, and are asked to annotate whether or not a claim is supported by the other text. This gives human annotations for INFOF1-Ref. (3) **Grounding Judgments**. Human annotators are given all the relevant videos for a topic and the predicted subclaims and asked to annotate whether or not the subclaims are supported by (grounded in) the videos. This gives human annotations for CITEP.

In the following paragraphs, we will summarize the agreements and findings between human judgments and automatic methods, but leave discussion on *how* to evaluate multimodal RAG to section 6. The agreement is a Kendall's Tau (Kendall, 1938) between the human and metric ranking of systems.

Human MIRAGE In Table 2, we explore the agreement between the extrinsic quality judgments (EQJs) and the two metric specific quality judgments: intrinsic claim judgments (ICJs, mirroring INFOF1-Ref) and grounding judgments (GJs, mirroring CITEP). We find INFOF1 *aligns strongly with human annotated judgments* for the properties considered in the EQJs, higher than n-gram (ROUGE) and well calibrated semantic similarity (BERTScore) metrics. This provides evidence that with well-calibrated systems of INFOF1, MIRAGE *is a strong metric for both interpretable information judgments and broad quality judgments*.

We also observe that grounding judgments (GJs) do not align with EQJs. This is not unexpected because the EQJs only evaluate quality with the reference and no qualities related to the actual videos. However, this result is important showing that *metrics that focus on capturing EQJs or ICJs may not align with GJs unless they verify against the sources of information*.

Automatic Information Scores In Table 1, we report the agrement of each automatic metric variant that evaluates information—INFOF1 (INFOP, INFOR), ALCE Claim Recall (Clm. Rec.), ARGUE Nugget Coverage (Nugg. Cov.), RAGAS Faithfulness (Faith.) and Answer Relevance (Ans. Rel), ROUGE, and BERTScore—with the three human annotated judgments. We find that ROUGE

[5] See Appendix G for instructions and ranking.

CITEP		CITER	AC	ARG	RG
R	C	R	CQ	SS	CR
11.5	42.7	54.3	20.0	-15.0	32.7
11.3	49.7	55.0	10.0	1.8	23.2
21.5	57.7	57.7	8.2	-1.8	13.2
44.8	16.7	4.8	8.2	-3.3	10.0

Table 3: Kendall's Tau with for all human judgments against the metrics that evaluate citations. Row 1: EQJ, Row 2: GJ. AC: ALCE, ARG: ARGUE, RG: RAGAS

and BERTScore outperform all metrics when evaluating quality (EQJs) and the information in the predictions (EQJ/ICJ), but *fail to capture how grounded predictions are in their sources* (GJ). Additionally, we find claim-based precision metrics, RAGAS Faith and INFOP, not only align with EQJs and ICJs, but also with GJs. This demonstrates the *benefit of verifying information against the sources of information as a factuality metric*.

We note that the agreement across the board for these automatic metrics is low, outside of ROUGE and BERTScore. We attribute this to the fact that VLMs and LLMs are not calibrated for support judgments,[6] leading to lower agreement with the human judgments than human annotated metrics.

Automatic Citation Scores In Table 3, we report the alignment of each automatic metric variant that evaluates citations—CITEF1 (CITEP, CITER), ALCE Citation Quality (CQ), ARGUE Sentence Support (SS), and RAGAS Context Relevance (CR)—with the three human annotated judgments. We find that the metrics that evaluate sentences against sources do not agree with EQJs or GJs, having low or negative agreement. This demonstrates the claim that *evaluating against sentences negatively impacts support judgments*. Additionally, we find that CITEF1 has the highest agreement for both EQJs and GJs, validating the *benefits of subclaim decomposition in support judgments*.

Like the information scores, we note that the agreement between citation metrics and human judgments can be increased with the calibration of VLMs and LLMs for support judgments.

6 How to Evaluate Multimodal RAG

Evaluating any task is challenging. If there was a correct way to evaluate RAG, we would only need one metric. We offer two positions for evaluating multimodal RAG for quick diagnostics, and comprehensive system evaluation. **When we recommend reporting INFOF1 and CITEF1, we *always* recommend reporting Precision and Recall as they are fundamental, the F1 is derived.**

Quick or Constrained As mentioned above, the evaluation of predictions against multimodal sources can be computationally expensive, leading to significantly higher wall-clock time than the reference-based metrics. So, for performing quick diagnostics or evaluating in a compute constrained setting, we recommend evaluating multimodal RAG with ROUGE, BERTScore, INFOF1 (ref), and CITEF1 (ref). These metrics provide an effective balance between efficiency and interpretability: ROUGE and BERTScore offer fast, surface-level checks; INFOF1 and CITEF1 extend this with claim level reasoning, while avoiding the cost of multimodal inference. Together, they enable a lightweight evaluation of system behavior, ideal for iterative development or rapid benchmarking under limited compute budgets.

Comprehensive Evaluation When performing a comprehensive evaluation, it is important to assess predictions against the sources rather than the references. We recommend evaluating with ROUGE, BERTScore, INFOF1 (col), CITEF1 (col), and RAGAS-Faithfulness. This combination of surface-level, claim-level, and citation-level metrics provides a holistic, yet granular, view of system quality, capturing not only the quality of information but also how accurately it is grounded in the retrieved sources.

7 Conclusion

We introduce MIRAGE, an evaluation framework for multimodal retrieval augmented generation. MIRAGE, consisting of INFOF1 and CITEF1, assesses predictions at the subclaim level in the factuality, information coverage, citation support, and citation attribution. Comparing to summarization metrics and our multimodal implementations of TextRAG metrics, we find that summariation metrics don't capture grounding and TextRAG metrics don't align with human preferences. Additionally, in our analysis of the automatic implementations, we find automatic judgments for claim support inaccurate and inefficient for multimodal and multi-source verification, suggesting interesting future work in multi-video inference.

[6] See Appendix H for claim verification performance.

Limitations

Fine-grained Entailment Judgments In our work, we only explore binary entailment judgments (supported, not supported). It has been shown that scalar judgments and uncertainty-aware evaluation provides more granular measures of model capability (Chen et al., 2020; Cheng et al., 2024; Jiang et al., 2024a; Yuan et al., 2024). Text-based calibration methods exist for enabling LLMs to make these judgments (e.g. Jiang et al., 2022; Jurayj et al., 2025; Wang et al., 2025), and the calibration of VLMs for these judgments provides ample space for interesting future work.

Multi-Video Inference Constraints As mentioned, the inability to provide multi-video inference limits the capabilities of each method presented. Extreme reduction of video frame rates to enable multi-video inference limits the performance of the VLM verifier and performing only single video judgments limits the ability to assess any cross-source inferences made by the generation system. Training VLMs for multi-video inference and optimizing multimodal inference are much needed future work for more holistic (multi-video) and scalable (optimization) evaluation.

References

Samuel Amouyal, Tomer Wolfson, Ohad Rubin, Ori Yoran, Jonathan Herzig, and Jonathan Berant. 2023. QAMPARI: A benchmark for open-domain questions with many answers. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 97–110, Singapore. Association for Computational Linguistics.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Samuel Barham, Orion Weller, Michelle Yuan, Kenton Murray, Mahsa Yarmohammadi, Zhengping Jiang, Siddharth Vashishtha, Alexander Martin, Anqi Liu, Aaron Steven White, Jordan Boyd-Graber, and Benjamin Van Durme. 2023. Megawika: Millions of reports and their sources across 50 diverse languages. *Preprint*, arXiv:2307.07049.

Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. 2020. Uncertain natural language inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8772–8779, Online. Association for Computational Linguistics.

Yunmo Chen, William Gantt, Tongfei Chen, Aaron White, and Benjamin Van Durme. 2023. A unified view of evaluation metrics for structured prediction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12868–12882, Singapore. Association for Computational Linguistics.

Qi Cheng, Michael Boratko, Pranay Kumar Yelugam, Tim O'Gorman, Nalini Singh, Andrew McCallum, and Xiang Li. 2024. Every answer matters: Evaluating commonsense with probabilistic measures. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 493–506, Bangkok, Thailand. Association for Computational Linguistics.

Xinya Du, Alexander Rush, and Claire Cardie. 2021. GRIT: Generative role-filler transformers for document-level event entity extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 634–644, Online. Association for Computational Linguistics.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

William Gantt, Alexander Martin, Pavlo Kuchmiichuk, and Aaron Steven White. 2024. Event-keyed summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7333–7345, Miami, Florida, USA. Association for Computational Linguistics.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Yanjun Gao, Chen Sun, and Rebecca J. Passonneau. 2019. Automated pyramid summarization evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 404–418, Hong Kong, China. Association for Computational Linguistics.

Anisha Gunjal and Greg Durrett. 2024. Molecular facts: Desiderata for decontextualization in LLM fact verification. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3751–3768, Miami, Florida, USA. Association for Computational Linguistics.

Deepak Gupta, Dina Demner-Fushman, William Hersh, Steven Bedrick, and Kirk Roberts. 2024. Overview of trec 2024 biomedical generative retrieval (biogen) track. *Preprint*, arXiv:2411.18069.

Rujun Han, Yuhao Zhang, Peng Qi, Yumo Xu, Jenyuan Wang, Lan Liu, William Yang Wang, Bonan Min, and Vittorio Castelli. 2024. RAG-QA arena: Evaluating domain robustness for long-form retrieval augmented question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4354–4374, Miami, Florida, USA. Association for Computational Linguistics.

Hang Hua, Yunlong Tang, Chenliang Xu, and Jiebo Luo. 2024. V2xum-llm: Cross-modal video summarization with temporal prompt instruction tuning. *Preprint*, arXiv:2404.12353.

Soyeong Jeong, Kangsan Kim, Jinheon Baek, and Sung Ju Hwang. 2025. Videorag: Retrieval-augmented generation over video corpus. *Preprint*, arXiv:2501.05874.

Zhengping Jiang, Anqi Liu, and Benjamin Van Durme. 2025. Conformal linguistic calibration: Trading-off between factuality and specificity. *Preprint*, arXiv:2502.19110.

Zhengping Jiang, Anqi Liu, and Benjamin Van Durme. 2022. Calibrating zero-shot cross-lingual (un-)structured predictions. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2648–2674, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhengping Jiang, Anqi Liu, and Benjamin Van Durme. 2024a. Addressing the binning problem in calibration assessment through scalar annotations. *Transactions of the Association for Computational Linguistics*, 12:120–136.

Zhengping Jiang, Jingyu Zhang, Nathaniel Weir, Seth Ebner, Miriam Wanner, Kate Sanders, Daniel Khashabi, Anqi Liu, and Benjamin Van Durme. 2024b. Core: Robust factual precision with informative sub-claim identification. *Preprint*, arXiv:2407.03572.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. 2024. FaithScore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5042–5063, Miami, Florida, USA. Association for Computational Linguistics.

William Jurayj, Jeffrey Cheng, and Benjamin Van Durme. 2025. Is that your final answer? test-time scaling improves selective question answering. *Preprint*, arXiv:2502.13962.

M. G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30:81–93.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *International Conference on Computer Vision (ICCV)*.

Reno Kriz, Kate Sanders, David Etter, Kenton Murray, Cameron Carpenter, Kelly Van Ochten, Hannah Recknor, Jimena Guallar-Blasco, Alexander Martin, Ronald Colaianni, Nolan King, Eugene Yang, and Benjamin Van Durme. 2025. Multivent 2.0: A massive multilingual benchmark for event-centric video retrieval. *Preprint*, arXiv:2410.11619.

Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W. Oard, Luca Soldaini, and Eugene Yang. 2024. Overview of the trec 2023 neuclir track. *Preprint*, arXiv:2404.08071.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. 2025. Videochat-flash: Hierarchical compression for long-context video modeling. *Preprint*, arXiv:2501.00574.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jingyang Lin, Hang Hua, Ming Chen, Yikang Li, Jenhao Hsiao, Chiuman Ho, and Jiebo Luo. 2024. Videoxum: Cross-modal visual and textural summarization of videos. *Preprint*, arXiv:2303.12060.

Jingyang Lin, Jialian Wu, Ximeng Sun, Ze Wang, Jiang Liu, Yusheng Su, Xiaodong Yu, Hao Chen, Jiebo Luo, Zicheng Liu, and Emad Barsoum. 2025. Unleashing hour-scale video training for long video-language understanding. *Preprint*, arXiv:2506.05332.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jingzhou Liu, Wenhu Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang, and Jingjing Liu. 2020. Violin: A large-scale dataset for video-and-language inference. *Preprint*, arXiv:2003.11618.

Xueguang Ma, Luyu Gao, Shengyao Zhuang, Jiaqi Samantha Zhan, Jamie Callan, and Jimmy Lin. 2025. Tevatron 2.0: Unified document retrieval toolkit across scale, language, and modality. *Preprint*, arXiv:2505.02466.

Alexander Martin, Reno Kriz, William Gantt Walden, Kate Sanders, Hannah Recknor, Eugene Yang, Francis Ferraro, and Benjamin Van Durme. 2025. Wikivideo: Article generation from multiple videos. *Preprint*, arXiv:2504.00939.

James Mayfield, Eugene Yang, Dawn Lawrie, Sean MacAvaney, Paul McNamee, Douglas W. Oard, Luca Soldaini, Ian Soboroff, Orion Weller, Efsun Kayi, Kate Sanders, Marc Mason, and Noah Hibbler. 2024. On the evaluation of machine-generated reports. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 1904–1915, New York, NY, USA. Association for Computing Machinery.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang 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*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Arun Reddy, Alexander Martin, Eugene Yang, Andrew Yates, Kate Sanders, Kenton Murray, Reno Kriz, Celso M. de Melo, Benjamin Van Durme, and Rama Chellappa. 2025. Video-colbert: Contextualized late interaction for text-to-video retrieval. *Preprint*, arXiv:2503.19009.

Xubin Ren, Lingrui Xu, Long Xia, Shuaiqiang Wang, Dawei Yin, and Chao Huang. 2025. Videorag: Retrieval-augmented generation with extreme long-context videos. *Preprint*, arXiv:2502.01549.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive learning with hard negative samples. *Preprint*, arXiv:2010.04592.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Saron Samuel, Dan DeGenaro, Jimena Guallar-Blasco, Kate Sanders, Seun Eisape, Arun Reddy, Alexander Martin, Andrew Yates, Eugene Yang, Cameron Carpenter, David Etter, Efsun Kayi, Matthew Wiesner, Kenton Murray, and Reno Kriz. 2025. Mmmorrf: Multimodal multilingual modularized reciprocal rank fusion. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 4004–4009, New York, NY, USA. Association for Computing Machinery.

Kate Sanders, Reno Kriz, David Etter, Hannah Recknor, Alexander Martin, Cameron Carpenter, Jingyang Lin, and Benjamin Van Durme. 2024. Grounding partially-defined events in multimodal data. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15905–15927, Miami, Florida, USA. Association for Computational Linguistics.

Ori Shapira, Ramakanth Pasunuru, Mohit Bansal, Ido Dagan, and Yael Amsterdamer. 2022. Interactive query-assisted summarization via deep reinforcement learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2551–2568, Seattle, United States. Association for Computational Linguistics.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024. VeriScore: Evaluating the factuality of verifiable claims in long-form text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9447–9474, Miami, Florida, USA. Association for Computational Linguistics.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8273–8288, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nandan Thakur, Ronak Pradeep, Shivani Upadhyay, Daniel Campos, Nick Craswell, and Jimmy Lin. 2025. Support evaluation for the trec 2024 rag track: Comparing human versus llm judges. *Preprint*, arXiv:2504.15205.

Siddharth Vashishtha, Alexander Martin, William Gantt, Benjamin Van Durme, and Aaron White. 2024. FAMuS: Frames across multiple sources. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8250–8273, Mexico City, Mexico. Association for Computational Linguistics.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, and 16 others. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Ellen M. Voorhees. 2004. Overview of the trec 2003 question answering track. In *Text Retrieval Conference*.

William Walden, Pavlo Kuchmiichuk, Alexander Martin, Chihsheng Jin, Angela Cao, Claire Sun, Curisia Allen, and Aaron White. 2025a. Cross-document event-keyed summarization. In *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)*, pages 218–241, Vienna, Austria. Association for Computational Linguistics.

William Walden, Marc Mason, Orion Weller, Laura Dietz, Hannah Recknor, Bryan Li, Gabrielle Kaili-May Liu, Yu Hou, James Mayfield, and Eugene Yang. 2025b. Auto-argue: Llm-based report generation evaluation. *Preprint*, arXiv:2509.26184.

Liaoyaqi Wang, Zhengping Jiang, Anqi Liu, and Benjamin Van Durme. 2025. Always tell me the odds: Fine-grained conditional probability estimation. *Preprint*, arXiv:2505.01595.

Miriam Wanner, Benjamin Van Durme, and Mark Dredze. 2024a. Dndscore: Decontextualization and decomposition for factuality verification in long-form text generation. *Preprint*, arXiv:2412.13175.

Miriam Wanner, Seth Ebner, Zhengping Jiang, Mark Dredze, and Benjamin Van Durme. 2024b. A closer look at claim decomposition. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 153–175, Mexico City, Mexico. Association for Computational Linguistics.

Yumo Xu and Mirella Lapata. 2021. Generating query focused summaries from query-free resources. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6096–6109, Online. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Moy Yuan, Eric Chamoun, Rami Aly, Chenxi Whitehouse, and Andreas Vlachos. 2024. PRobELM: Plausibility ranking evaluation for language models. In *First Conference on Language Modeling*.

Emmanouil Zaranis, António Farinhas, Saul Santos, Beatriz Canaverde, Miguel Moura Ramos, Aditya K Surikuchi, André Viveiros, Baohao Liao, Elena Bueno-Benito, Nithin Sivakumaran, Pavlo Vasylenko, Shoubin Yu, Sonal Sannigrahi, Wafaa Mohammed, Ben Peters, Danae Sánchez Villegas, Elias Stengel-Eskin, Giuseppe Attanasio, Jaehong Yoon, and 12 others. 2025. Movie facts and fibs (mf^2): A benchmark for long movie understanding. *Preprint*, arXiv:2506.06275.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. Optimizing dense retrieval model training with hard negatives. *Preprint*, arXiv:2104.08051.

Shiyue Zhang, David Wan, Arie Cattan, Ayal Klein, Ido Dagan, and Mohit Bansal. 2024. Qapyramid: Fine-grained evaluation of content selection for text summarization. *Preprint*, arXiv:2412.07096.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

A Appendix Road Map

Metrics Details You can find the variations of INFOF1 and CITEF1 in Appendix B. The details for the implementations of ALCE, ARGUE, and RAGAS can be found in Appendix D, Appendix E, and Appendix F, respectively.

Reference vs Collection For a discussion on reference vs collection precision for INFOF1 and CITEF1 see Appendix C.

Human Judgments In Appendix G you can find the tables showing the alignment between our human annotated metric, annotation instructions for each human judgment, and information on how the

rankings for human annotation and metric outputs were created.

Analysis of Verification Methods In Appendix H, you can see experiments and results for a naive baseline for calibrating LLMs and VLMs for claim verification in documents and videos.

Evaluating X-to-Text Appendix I includes a further discussion on other tasks where MIRAGE can be applied (e.g., summarization).

Related Work For an extended relaetd works section, see Appendix J

B MIRAGE in Detail

MIRAGE is a claim-centric evaluation framework with the aim of evaluating two main aspects: (1) INFOF1, information presented in a prediction is factual (precision) and fulfills the information need (recall), and (2) CITEF1, cited sources are relevant to the claims they are intended to support (precision) and recalled information cites proper sources (recall). In the following sections, we provide formulations for evaluation as presented in section 4 as well as variations for single-video inference and weighted variations for the importance of information and quality of sources.

B.1 INFOF1

In this section, we will start from the original formulation of INFOF1 and then reformulate precision and recall for single source inference and weighted variations.

Importance of Information In the weighted variations of INFOF1, we will use a notion of *information importance* to weight both precision and recall. For **precision** information importance should be defined by the author of the prediction. For example, a human could annotate a ranked list for the claims they present from most to least important. A model could do the same, or, for a calibrated model (e.g., Jiang et al., 2025), express confidence in the predicted claims as their importance. For **recall**, information importance should be defined by the evaluation data or the consumer of the information. In practice, this could be annotating each reference claim in a ranked list for it's importance to the queried information need.

B.1.1 Information Factuality (Precision)

To capture the factuality of a prediction (P), we look to evaluate the factuality of predicted subclaims against some existing knowledge on the topic ($E = [d_1, \ldots, d_n]$, an evidence base). We first decompose each predicted sentence into a set of subclaims (C_P). Then, these subclaims are scored for their support in the evidence using the scoring function. Formally,

$$\text{precision}(C_P, E) = \frac{1}{|C_P|} \sum_{c \in C_P} \mathtt{s}(c, E) \quad (6)$$

However, Equation 8 has two variations: (1) a single source inference option, where a claim is scored against a single item instead of the full E, and (2) a weighted variation, where a claim may have an importance.

Single Source Precision Single source precision is the more computationally feasible counterpart to the multi-source formulation. Instead of scoring a claim against all evidence, we can instead compare the claim individually against each item in the evidence, and take the maximum similarity as the score for the claim. Formally,

$$\text{precision}(C_P, E) = \frac{1}{|C_P|} \sum_{c \in C_P} \max_{d \in E} \mathtt{s}(c, d) \quad (7)$$

where we use d to denote the canonical term "document," but meaning any source of information.

While single source precision is more computationally feasible, it is limited by the inability to evaluate any "cross-source" inferences made by the prediction system during generation.

Weighted Precision For weighted precision, we define I_i as the importance for a claim c_i and define precision as follows:

$$\text{precision}(C_P, E) = \frac{1}{|C_P|} \sum_{c \in C_P} \mathtt{s}(c, E) * I \quad (8)$$

For single source inference this would be expressed as

$$\text{precision}(C_P, E) = \frac{1}{|C_P|} \sum_{c \in C_P} \max_{d \in E} \mathtt{s}(c, d) * I \quad (9)$$

B.1.2 Information Coverage (Recall)

To ensure that a prediction covers the relevant information from the evidence, we look to evaluate the coverage of reference subclaims against the prediction. We first decompose each reference sentence into a set of subclaims (C_R). Then, these subclaims

Eval	Ano	INFOF1		Precision		Recall
		Reference	Collection	Reference	Collection	
EQJ	1	34.5	23.3	26.3	23.3	16.3
	2	41.2	29.3	18.2	31.0	31.2
	3	28.2	28.2	17.8	28.2	13.0
ICJ	1	35.2	19.7	38.2	11.5	3.5
	2	40.0	36.3	26.7	16.3	16.7
	3	38.2	27.8	26.7	14.8	12.2
GJ	NA	6.6	6.6	6.6	21.5	1.8

Table 4: Kendall's Tau for the components of INFOF1. For Precision and Recall judgments against ICJ, the score is computed against the human equivalent, not the human f1. EQJ: Extrinsic Quality Judgment, ICJ: Intrinsic Claim Judgment, GJ: Grounding Judgment

are scored for their support in the prediction (P) by the scoring function. Formally,

$$\text{recall}(C_R) = \frac{1}{|C_R|} \sum_{c \in C_R} \mathsf{s}(c, P) \quad (10)$$

Weighted Precision Like it's counter part, INFOF1 Precision, this can also be weighted by an importance term.

$$\text{recall}(C_R) = \frac{1}{|C_R|} \sum_{c \in C_R} \mathsf{s}(c, P) * I \quad (11)$$

B.1.3 Granular Human Alignment

In the main text, we presented how the F1 versions of INFOF1 aligned with human judgments. However, we can perform the same alignment calculations using Precision and Recall. In Table 4, we evaluate the agreement between INFOF1 precision and recall components and the human judgments. Note, for ICJ, precision and recall are compared with their respective human annotated components and not the human annotated F1. We have two main takeaways from these results.

1. We find a large discrepancy between support annotated by a zero-shot LLM and zero-shot VLM for INFOF1 precision, with agreement significantly dropping when annotating support with a VLM.

2. We find that LLMs struggle with recall judgments.

In Appendix H, we explore the performance of the underlying LLM and VLM on verifying claims in videos or documents, but broadly note the calibration of these models as an interesting avenue for future work.

Paper Implementation For INFOF1, we implement the single-source variation in our paper, prompting the verifiers with Figure 3 or Figure 4 for the VLM and LLM, respectively.

B.2 CITEF1

In this section, we wills tart from the original formulation of CITEF1 and then reformulate precision and recall for single source inference and weighted variations.

Quality of Sources In the weighted variations of CITEF1, we will use a notion of *source quality* to weight both precision and recall. For both **precision** and **recall** the quality of a source could be annotated in the evaluation data. There are two more obvious cases of when you might want to rate the quality of a source: (1) differentiating between counterfactual sources (e.g., AIGC or misinformation), and (2) based on human preferences. For example, a human might prefer that a system cite the *most informative* sources, covering the most information, in order to have to read/view/watch fewer information sources to validate information presented in a prediction.

B.2.1 Citation Support (Precision)

To measure if the predicted citations (D_P) support the prediction, we look to evaluate the support between predicted subclaims and their associated citations. We first decompose subclaims and create a mapping between the subclaims and the videos cited by the sentences containing the claims ($c_i : \{v_1, \ldots, v_n\}$. This set of documents ($D_P(c_i)$) can either be the videos associated with the first mention c_i, like ARGUE, or any mention of c_i depending on the desired evaluation setting. However,

Eval	Ano	CITEF1 Reference	CITEF1 Collection	Precision Reference	Precision Collection	Recall
EQJ	1	55.7	57.5	11.5	42.7	54.3
	2	61.5	58.2	11.3	49.7	55.0
	3	59.0	60.8	21.5	57.7	57.7
ICJ	1	63.0	61.2	-6.6	20.0	79.3
	2	64.8	63.0	1.8	23.3	74.5
	3	68.2	66.3	1.8	36.7	76.0
GJ	NA	4.8	8.2	44.8	16.7	4.8

Table 5: Kendall's Tau for the components of CITEF1. For Precision and Recall judgments against ICJ, the score is computed against the human equivalent, not the human f1. EQJ: Extrinsic Quality Judgment, ICJ: Intrinsic Claim Judgment, GJ: Grounding Judgment

claim de-duplication is recommended as shown by Jiang et al. (2024b). Precision can be expressed as

$$\text{precision}(P, D_P) = \frac{1}{|C_P|} \sum_{c_i \in C_P} \mathtt{s}(c_i, D_P(c_i)) \tag{12}$$

B.2.2 Proper Attribution (Recall)

To determine if the proper sources are being attributed by the prediction, we look to evaluate if each reference subclaim covered in the prediction is attributed to a source that supports the subclaim. We first decompose the reference and create a mapping between each subclaim in the reference and the videos which could be cited to support this subclaim ($c_i^R : \{v_1, \ldots, v_n\}$), denoted cv_i for the i-th subclaim. Next, we create another mapping, this time using the prediction, between each cited video and the sentences that cite the video ($v_i : \{s_1^P, \ldots, s_m^P\}$), denoted vs_i for the i-th cited video. Using these two mappings, we then evaluate if a reference claim is recalled by the set of predicted sentences that cite a video which contains a reference claim. We take the maximum support for a reference claim from the set of possibly cited videos. Formally,

$$\text{recall} = \frac{1}{|C_R|} \sum_{c_i \in C_R} \max_{cv_i \in CV} \mathtt{s}(c_i, vs_{P_i}) \tag{13}$$

B.3 Granular Human Alignmnet

In the main text, we presented how the F1 versions of CITEF1 aligned with human judgments. However we can perform the same alignment calculations using precision and recall. In Table 5, we evaluate the agreement between CITEF1 precision and recall components and the human judgments. Note, for ICJ, precision and recall are compared with their respective human annotated components and not the human annotated F1. We have two main takeaways from these results.

1. We interestingly find that Collection Precision has higher correlation with extrinsic quality judgments than the reference version. We believe this is due to the descriptive nature of VLMs, which often produce subclaims like "smoke billowed from the tower." We discuss this more in Appendix C, but this low agreement is broadly due to the reference precision penalizing visual descriptions that are grounded in videos, but not in the high-level references.

2. We note that the low agreement between ICJ precision and CITEF1 precision is expected. There's a difference in a claim being supported in a video "smoke billowed from the tower" and what a human would include in an article. However, high recall is expected because it matters if the information is recalled from the reference to then be matched to the same citation.

3. We find that the only metric citation metric that correlates with human grounding judgments is INFOF1 Precision.

Paper Implementation For CITEF1, we implement the single-source variation in our paper, prompting the verifiers with Figure 3 for the VLM and either Figure 4 (precision) or Figure 5 (recall) for the LLM.

System Prompt: You are an expert in evaluating and verifying claims. You will be given a video, a claim, and the context the claim came from. Your task is to determine if the claim is supported by the video. You will output <response>yes<response> if the claim is supported by the video, or <response>no<response> if the claim is not supported by the video.
User Prompt: [VIDEO_HERE] Here is the context the claim came from: <claim_context> [PUT_CONTEXT_HERE] <claim_context>. Here is the claim: <claim> [PUT_CLAIM_HERE] <claim>. Only respond with <response>yes<response> or <response>no<response>. Is the claim: [PUT_CLAIM_HERE], supported by the video?

Figure 3: Zero-Shot Prompt for Claim Verification in Videos

System Prompt: You are an expert in evaluating and verifying claims. You will be given a passage of text, a claim, and the context the claim came from. Your task is to determine if the claim is supported by the passage of text. You will output <response>yes<response> if the claim is supported by the passage, or <response>no<response> if the claim is not supported by the passage.
User Prompt: Here is the passage: <verification_context> [PUT_VERIFICATION_CONTEXT_HERE] <verification_context>. Here is the context the claim came from: <claim_context> [PUT_CONTEXT_HERE] <claim_context>. Here is the claim: <claim> [PUT_CLAIM_HERE] <claim> Only respond with <response>yes<response> or <response>no<response>. Is the claim: [PUT_CLAIM_HERE], supported by the passage?

Figure 4: Zero-Shot Prompt for Claim Verification in Text

System Prompt: You are an expert in evaluating and verifying claims. You will be given a a claim, the context the claim came from, and a list of claims to verify the claim against. Your task is to determine if the claim is supported by the list of claims. You will output <response>yes<response> if the claim is supported by the list of claims, or <response>no<response> if the claim is not supported by the list of claims.
User Prompt: Here is the list of claims to verfiy against: <verification_context> [PUT_VERIFICATION_CONTEXT_HERE] <verification_context>. Here is the context the claim came from: <claim_context> [PUT_CONTEXT_HERE] <claim_context>. Here is the claim: <claim> [PUT_CLAIM_HERE] <claim> Only respond with <response>yes<response> or <response>no<response>. Is the claim: [PUT_CLAIM_HERE], supported by list of claims to verify against?

Figure 5: Zero-Shot Prompt for Claim Verification in Citation (Text)

C Reference vs Collection Metrics

In this section, we will discuss the tradeoffs between using reference vs collection precision in INFOF1 and CITEF1. We will primarily discuss the side effects of using reference vs collection precision, but note that the primary reason for using reference-based evaluations in multimodal RAG evaluation is for computational efficiency.

C.1 INFOF1 Reference and Collection Precision

To reiterate the main text, reference precision for INFOF1 evaluates predicted subclaims against the reference instead of against the videos related to the topic (collection precision). However, human annotated references are inherently non-exhaustive for all the possible information that could be included in a prediction. This leads to the artifact that **reference precision penalizes factual claims that are deemed not salient to the topic**.

This artifact may be a limitation if you are looking to evaluate *only* factuality of the prediction. Without judging the subclaim against the information sources, the evaluation cannot capture the factuality of *all* claims in the prediction. We do not view this artifact as a limitation for general evaluation because it is important that information be relevant to the query/information need.

C.2 CITEF1 Reference and Collection Precision

To reiterate the main text, reference precision for CITEF1 evaluates predicted subclaims against the reference subclaims that are grounded in the same cited video. However, like INFOF1 precision, human annotated references are non-exhaustive for all the possible information that could be included in a prediction. This leads to the artifact that **reference precision penalizes factual claims that are not deemed salient to the source topic**.

Unlike INFOF1 precision, this **is a limitation** of reference precision. CITEF1 is made to purely evaluate grounding in the cited sources. It does not matter if the information is not relevant, only that it is supported by the attributed sources.

D ALCE in Detail

ALCE (Gao et al., 2023) is a RAG evaluation framework developed for question answering with citations—the task of producing short-form (Stelmakh et al., 2022; Amouyal et al., 2023) or long-form (Fan et al., 2019) answers with citations attributing the answer source. ALCE introduces a correctness evaluation and citation evaluation.

Evaluation of Correctness When evaluating open-ended RAG on ELI5 (Fan et al., 2019), Gao et al. (2023) present Claim Recall, an NLI based metric between subclaims in the reference (gold answers) and the prediction (model output). This transfers to VideoRAG with no modification and is similar to our INFOF1 Recall.

In their paper, Gao et al. (2023) also introduce precision, but their method is exact match based and intended only for the short-form QA tasks (Stelmakh et al., 2022; Amouyal et al., 2023). This was not reasonable to use, nor did Gao et al. (2023), for evaluating long-form outputs.

Citation Quality Gao et al. (2023) evaluate citation quality with an F1 metric of citation recall and precision. When evaluating citations, Gao et al. (2023) compare a sentence from the prediction to the concatenation of cited sources. This poses two main issues in evaluation: (1) from a factuality perspective, evaluating at the sentence level requires veryfing multiple subclaims in a single inference step (Min et al., 2023), and (2) from a multimodal perspective, concatenating multiple videos raises issues with computational constraints and the non-existence of multiple video training data for VLMs (Li et al., 2025; Martin et al., 2025).

To faithfully implement citation quality, we wrap verification in a loop which downsamples the framerate of the video(s) until it fits on GPU memory. While this provides a faithful re-implementation of citation quality presented by Gao et al. (2023), the extreme downsampling *will* degrade performance.

Granular Human Alignment In Table 6, we evaluate the precision and recall breakdown of the citation quality metric.

Paper Implementation To **downsample multiple videos**, we first attempt to verify a (subclaim, videos) pair at 1 fps per video. If an out-of-memory (OOM) exception is thrown by the process, the videos are downsampled by reducing the framerate of all videos by half ($\frac{\text{fps}}{2}$) until 10 iterations of downsampling or until no OOM is reached. We never reach the 10 iterations on the videos in WikiVideo, but allow for setting downsampling iterations, to be defined by the evaluator. To verify the claims we use the prompt in Figure 6.

Eval	Ano	Claim Recall	Citation Quality F1	Citation Quality Precision	Citation Quality Recall
EQJ	1	-5.0	20.0	11.8	13.2
	2	21.2	10.0	5.0	1.7
	3	5.0	8.2	8.2	-13.3
ICJ	1	27.8	8.2	0.0	-6.6
	2	13.3	8.2	16.3	-1.8
	3	23.0	8.2	0.0	-6.6
GJ	NA	-8.2	8.2	0.0	6.6

Table 6: Kendall's Tau for the components of ALCE. For Precision and Recall judgments against ICJ, the score is computed against the human equivalent, not the human f1. EQJ: Extrinsic Quality Judgment, ICJ: Intrinsic Claim Judgment, GJ: Grounding Judgment

```
System Prompt: You are an expert at verifying information. You will be given a set of videos and a
sentence. Your task is to determine if the sentence is fully supported by the videos. You will output
<response>yes<response> if the sentence is fully supported by the videos, or <response>no<response>
if the sentence is not fully supported by the videos.
User Prompt: [VIDEOS_HERE] Sentence: [PUT_SENTENCE_HERE] Is the sentence fully supported by the
videos? Only respond with <response>yes<response> or <response>no<response>."
```

Figure 6: Zero-Shot Prompt for ALCE Citation Quality

E ARGUE in Detail

ARGUE (Mayfield et al., 2024) is a RAG evaluation framework developed for *report generation*—the task of producing a long-form, citation-attributed response to a complex user query. Under ARGUE, reports accumulate a series of rewards and penalties based on sentence-level judgments about whether each sentence is supported by its associated citations and whether it correctly answers any of a predetermined set of questions (or *nuggets*) associated with the report topic. ARGUE also handles various nuances related to these two dimensions, including whether a sentence *reiterates* information stated previously in the report (and thus does not require a citation) or whether a statement correctly asserts that some nugget question is unanswerable from the underlying collection.

Nuggets To evaluate predictions, ARGUE has a knowledge base of question-answer pairs on the topic called nuggets. We convert our claims to nuggets by generating questions with an LLM (Qwen3; Yang et al., 2025) and then pairing the question and claim as a question-answer pair.

Sentence Precision is the proportion of predicted sentences attested by each of their citations. The open-source implementation of ARGUE (Walden et al., 2025b) we use implements this for a single citation at a time, which we follow. However, this sentence precision still faces the same issue as ALCE from a factuality perspective.

Nugget Recall is the proportion of nugget questions correctly answered by the report, aggregating over sentences. This is similar to ALCE Claim Recall or our INFOF1 Recall, but instead recalling QA pairs. Like, ALCE this score requires no modification for multimodal RAG evaluation.

Granular Human Alignment In Table 7, we evaluate the precision and recall breakdown of ARGUE.

Paper Implementation Our implementation of ARGUE builds largely off of the automatic implementation from Walden et al. (2025b), with the only variation being the change of sentence support to evaluate a sentence against a video.

F RAGAS in Detail

RAGAS (Es et al., 2024) is a RAG evaluation framework developed for the automatic evaluation of retrieval-augmented generation–the task of producing long-form, citation attriburted responses to a user query. RAGAS consists of three main components: Faithfulness, Answer Relevance, and Context Relevance.

Eval	Ano	ARGUE	Sentence Support	Nugget Coverage
EQJ	1	-1.5	-3.5	-15.0
	2	1.8	33.0	1.8
	3	-1.8	11.3	-1.8
ICJ	1	10.0	13.2	26.7
	2	10.0	-4.8	38.3
	3	1.8	10.0	31.5
GJ	NA	-3.3	-3.3	4.8

Table 7: Kendall's Tau for the components of ARGUE. For Precision and Recall judgments against ICJ, the score is computed against the human equivalent, not the human f1. The ICJ judgments for Sentence Support and Nugget Coverage differ from the ~ text because of this. EQJ: Extrinsic Quality Judgment, ICJ: Intrinsic Claim Judgment, GJ: Grounding Judgment.

Eval	Ano	Faithfulness	Answer Relevance	Context Relevance
EQJ	1	32.8	4.8	32.7
	2	29.8	23.3	23.2
	3	16.2	-1.5	13.2
ICJ	1	-6.7	44.8	18.2
	2	8.5	33.3	24.8
	3	1.8	23.3	31.5
GJ	NA	24.8	-26.7	10.0

Table 8: Kendall's Tau for the components of RAGAS. For Precision and Recall judgments against ICJ, the score is computed against the human equivalent, not the human f1. The ICJ judgments for Faithfulness differ from the main text because of this. EQJ: Extrinsic Quality Judgment, ICJ: Intrinsic Claim Judgment, GJ: Grounding Judgment

Faithfulness evaluates "statements" from a prediction against the context used in generation. We interpret these statements as subclaims to resolve the factuality issues of ARGUE and ALCE. Additionally, like Citation Quality from ALCE, this experiences the same issue of computational infeasibility in multi-source inference. However, unlike ALCE, this is not designed such that it is necessary to perform inference on multiple sources. We modify this to verify one subclaim in one video at a time, thus becoming similar to our INFOF1 Collection Precision, but only using the videos used in generation and not all relevant videos.

Answer Relevance evaluates the embedding similarity between the query used during prediction and an LLM generated query based on the prediction. This metric remains unchanged from Es et al. (2024) because the generation and comparison of potential queries is text-based no matter the information source modality.

Context Relevance extracts sentences from the context based on the query and computes relevance as $\frac{\text{num extracted sentences}}{\text{total number of sentences}}$. We attempt to implement this with a VLM, extracting all information relevant to the query (numerator) and extracting all information from the video (denominator). However, extracting all information from the video is challenging, as the number of potential claims is near infinite. We instead elicit a detailed summary from a VLM for the denominator.

In the RAGAS implementation, Es et al. (2024) give the LLM the ability to backoff and say that a source has "insufficient information." We notice in our implementation of this that WikiVideo is challenging for VLMs to make high-level inferences about and thus leads to a large number of videos deemed "insufficient information," even though humans have annotated them and grounded relevant claims to the query topic in those same videos.

Granular Human Alignment RAGAS has no F1 score, so the only difference in Table 8 and the main text is that RAGAS-Faithfulness is evaluated against ICJ precision and not ICJ F1.

Paper Implementation We verify statements (subclaims) against the video content using the same prompt as MIRAGE for videos (Figure 3). For context relevance, we generate the potential query with the 7b version of Qwen3 and do the embedding similarity between the real query and predicted query with Qwen3 embed 0.6B. For context relevance, we extract information related to the query with Figure 7 (numerator) and detailed video information with Figure 8 (denominator).

G Collection of Human Judgments

To evaluate the various metrics introduced above, we collect three human judgments.[7]

1. **Extrinsic Quality Judgments**. Human annotators are given information about the topic and are asked to consider the following attributes (in order of importance): factuality, adequacy, coherence, relevancy, and fluency.

2. **Intrinsic Claim Judgments**. Human annotators are given the prediction, predicted subclaims, reference, and reference subclaims, and are asked to annotate whether or not a claim is supported by the other text. **This gives human annotations for INFOF1 (ref).**

3. **Grounding Judgments**. Human annotators are given all the relevant videos for a topic and the predicted subclaims and asked to annotate whether or not the subclaims are supported by (grounded in) the videos.

For each grounding judgment, we calculate the ranking between the three prediction settings and we use the scipystats (Virtanen et al., 2020) implementation of Kendall's Tau (Kendall, 1938).

G.1 Extrinsic Quality Judgments

In Figure 9, we provide the annotation instructions for the extrinsic quality judgments. For these judgments, we recruit three annotators who are native/fluent English speakers and provide them each the 10 evaluation instances and annotation instructions for the task. These annotations are redundant.

G.2 Intrinsic Quality Judgments

In Figure 10, we provide the annotation instructions for intrinsic quality judgments. For these judgments, we recruit three annotators who are native/fluent English speakers and provide them each the 10 evaluation instances and annotation instructions for the task. These annotations are redundant.

[7] See Appendix G for all annotation instructions and human ranking calculations.

```
Please extract relevant sentences from the provided context that can potentially help answer the following question. If no relevant sentences are found, or if you believe the question cannot be answered from the given context, return the phrase 'Insufficient Information'. The question is: [PUT_QUESTION_HERE]
```

Figure 7: Zero-Shot prompt for extracting the detailed summary used in the denominator of RAGAS Context Relevance (numerator).

```
Describe the video in detail and extract all the information possible from it. This includes transcribing any on screen text (OCR) and describing any visual information beyond the summary.
```

Figure 8: Zero-Shot prompt for extracting the detailed summary used in the denominator of RAGAS Context Relevance (denominator).

G.3 Grounding Judgments

We use the same annotation instructions from Martin et al. (2025) and refer the read to their work for the annotation instructions. For these judgments, we recruit 5 annotators who are native/fluent English speakers and provide them the annotation instructions and the annotation data. These annotations are not redundant.

G.4 Human MIRAGE Granular Alignment

In Table 9 we provide the alignment between each component (Precision, Recall) of human annotated INFOF1 and extrinsic quality judgments.

H Usability of Verification Methods

In this section, we first discuss the performance of claim verification methods and then explain the training setting for each verification method.

H.1 Verifying Verifiers

To verify whether a subclaim is verified in a video or text, we use (V)LMs to judge the support between a claim and source pair. We specifically setup the verification to as $\mathtt{s}([\text{context}][\text{subclaim}], [\text{source}])$, where the context is the text from which the subclaim was decomposed and the source is the video or text that is being judged to support the subclaim. We provide the context to provide a form of decontextualization, as proposed by Wanner et al. (2024a), without the potential errors of decontextualizaton.

Are VLMs Calibrated for Verification? LLMs are known to not be calibrated for probablistic UNLI judgments (Chen et al., 2020; Wang et al., 2025), however, it is widely accepted they are adequate for binary (yes/no) judgments (Min et al., 2023). However, no work has explored the performance of VLMs in verifying (subclaim, video) pairs. To test this, we filter the subclaims of the evaluation topic to be balanced (there's a large class imbalance towards negative). In Table 10, we report the performance of Qwen2.5 VL in zero-shot (ZS) for the filtered claims on the evaluation topics, finding *VLMs are not calibrated for ZS verification*.

Training a Video-Claim Verifier Table 10 shows that the performance of claim verification by a VLM in a ZS is not adequate for automatic evaluation. To attempt to alleviate this issue, we train a VLM verifier on the set of WikiVideo topics **not** included in our human evaluation set (47 topics). We filter down the subclaims to be an equal distribution of supported/not and train on those 1988 subclaims. We report the performance on the evaluation in Table 10, finding that *this provides some calibration for the task in verification as next-token prediction (NT), but not for natural language inference (NLI)*.

Is Verifying in the Reference Equivalent? In both methods of reference precision, we verify in the reference article instead of the video content. Thus, we need to ensure that verifying claims in the reference performs well. For this, we first verify if LLMs are calibrated for the task and then follow a similar training to video-claim verifier. In Table 10, we find that *reference verification performs better than SFT VLM inference in ZS and after SFT*.

Note, we select the ZS variations, even if they underperform the SFT counterpart, for the main analysis in section 5 because the TextRag metrics get a ZS model for sentence support judgments.

H.2 Training a Verification Model

Architecture We adopt two complementary approaches: Natural Language Inference (NLI) and Next Token Prediction (NT) for verification.

For the NLI approach, we design classification heads for both the LLM and VLM backbones, in-

	ICJ 1			ICJ 2			ICJ 3		
	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
EQJ 1	49.0	37.8	40.8	65.3	68.7	59.0	60.5	36.0	52.3
EQJ 2	61.3	49.7	68.0	76.3	53.0	74.5	81.2	51.5	81.2
EQJ 3	54.5	43.3	64.2	70.8	57.8	77.2	72.7	49.7	74.2

Table 9: Kendall's Tau agreement between MIRAGE human annotated judgments (ICJ[1-3]) and human annotated quality judgments (EQJ[1-3]).

Modality	Method	F1	AUC
None	Yes	50.0	50.0
QVL2.5-NT	ZS-7B	41.7	53.2
	SFT-7B	76.9	75.4
QVL2.5-NLI	ZS-7B	35.6	55.2
	SFT-7B	52.7	53.4
Q2.5-NT	ZS-7B	83.1	85.6
	SFT-7B	99.4	99.4
Q2.5-NLI	ZS-7B	37.2	47.4
	SFT-7B	88.7	88.7

Table 10: Qwen2.5 (LLM: Q2.5 or VLM: QVL2.5) performance in claim verification under natural language inference (NLI) or next token (NT) settings. Underscored rows note the methods used in section 5.

stantiated with Qwen2.5-7B-Instruct(Qwen et al., 2025) and Qwen2.5-VL-7B(Bai et al., 2025), respectively. These heads output prediction probabilities, which are then aggregated to determine whether a given claim is supported by the corresponding video or article evidence.

For the NT approach, we leverage the generative capabilities of Qwen2.5 models by employing a structured prompting strategy. Specifically, the model is instructed to produce answers enclosed within `<response>...</response>` tags. The extracted responses are parsed and compared against the claim to assess supportiveness.

Data To construct the training and evaluation sets, we retain all instances with true labels and sample an equal number of false-labeled instances. For the negative examples, we incorporate hard negatives as part of the sampling process to enhance model performance. Hard negative mining is a well-established technique in information retrieval and classification, known to improve performance through more challenging training signals (Robinson et al., 2021; Zhan et al., 2021). Following this paradigm, we first classify all claims annotated by Martin et al. (2025) using Qwen2.5 VL 72B (Bai et al., 2025). We then construct our training set by combining 25% randomly sampled negatives with 75% hard negatives. This yields a balanced training set consisting of 1988 claims (supported vs. not supported), and an evaluation set of 787 claims.

Training In our training settings, both NLI and NT methods are trained for 10 epochs. The NLI heads are optimized with a learning rate of 2×10^{-4}, while NT training uses a smaller learning rate of 2×10^{-5} to better preserve generative quality.

I Evaluating X-to-Text

In this section, we will discuss how to use MIRAGE under common Text-to-Text, Image-to-Text, Audio-to-Text, Video-to-Text, and Any-to-Text settings. Instead of discussing the implementation at the modality level, we will instead discuss the implementation for single-source and multi-source generation, and then discuss what would be needed to use our evaluation framework in that setting.

Single-Source Inference When performing the evaluation of text generated from a single source, the formulation in Appendix B for single source inference is suitable for both precision metrics. Additionally, INFOP and CITEP become equivalent evaluation metrics, since each subclaim should be verified against the single source. There is no need for CITER because all information will be recalled from the same source.

Multi-Source Inference Following the paper will lead to succesful implementations for any other modality. In our work, we specifically evaluate against videos (containing both audio and visual signals). By extension the same models can be used for image and audio verification.

Implementation for other modalities As we noted in section 5, LLMs and VLMs need cali-

For each prediction, you should give a likert score from 1 to 5. The criteria for giving a score is as follows (in descending importance):

- **Consistency/Factuality:** Does the article make only true statements about the topic in question, given what the reference says about that topic?
  - Articles that make factual errors, omissions, or hallucinations of any kind should be penalized.
- **Adequacy:** Does the article adequately capture all of the information contained in the reference article?
  - Articles that omit details about any of the important aspects of the topic should be penalized.
- **Coherence:** Does the article make sense on its own, as a standalone description of the topic?
  - Articles that require you to go read the reference in order to understand what they mean (or that don't make sense even then) should be penalized.
  - Articles that don't provide substance on the event should also be penalized (e.g. "the [event] happened at [a time] in [a place]")
- **Relevancy:** Does the article include only information that is relevant to the topic in question?
  - Articles that include irrelevant or superfluous information, or information about some topic other than the one represented by the reference article, should be penalized.
- **Fluency:** Does the article sound reasonable natural (like something a native English speaker might actually write)?
  - Articles that are disfluent or that sound unnatural should be penalized.

Oftentimes, some of the summaries may be very similar to each other. It is totally fine to give multiple summaries the same score if you think they are of comparable quality!
You should enter your score for each summary in the score field. The default value for each summary is 0, meaning you have not yet annotated the likert judgment. Please do not use half scores (1.5, 2.5, etc).

Figure 9: Instructions for the human likert judgments (EQJ)

In this task, you'll be given a set of 3 json files for each prediction and 3 json files for the reference. The reference json files will all be the same, but each will correspond to your annotations for a different system's predictions. In each json file, there will be the same 10 topics.
For every system prediction json, you'll see

1. **prediction:** the predicted article for the topic. This prediction is the model's attempt to write an article for the topic.

2. **claims:** this is a sub dictionary that contains the claims and sentences they came from it is formatted below. The judgment is whether the claim is supported by the reference article or not.

```
{ "sentence": { "claim": "the claim to be verified", "judgment": [False|True] }, }
```

Annotation Task
The goal of this task is to mirror how we evaluate multimodal RAG in InfoF1. The basic premise is as follows.

- For every claim in the prediction, check if it is supported by the reference article. If it is mark the judgment as True, otherwise mark it as False. (The default value is None)

- For every claim in the reference, check if it is supported by the prediction article. If it is mark the judgment as True, otherwise mark it as False. (The default value is None)

Claim Annotation Criteria
When deciding whether a **claim is supported** by the other article (reference or prediction), apply the following criteria. The goal is to make judgments **only when there is no reasonable doubt** that the claim is supported by the other text.
A claim is **SUPPORTED (True)** if:

1. **All factual elements** in the claim are explicitly stated or can be directly inferred from the other article without needing external knowledge.
   - Example: Claim: "The Eiffel Tower is located in Paris." → Supported if the other article states "The Eiffel Tower is in Paris."

2. The **meaning and intent** of the claim are **fully consistent** with the information in the article.
   - Minor wording differences are acceptable if they don't change meaning.

3. The **temporal or causal context** matches (e.g., dates, events, outcomes are consistent).

4. If the claim includes **quantitative or categorical facts** (numbers, names, locations, affiliations), these details must exactly match what is stated in the other article.

A claim is **NOT SUPPORTED (False)** if:

1. The other article **contradicts** any part of the claim.

2. The other article **omits or is ambiguous** about key details needed to verify the claim.

3. The claim requires **inference beyond what's stated**, such as outside knowledge, assumptions, or general reasoning not grounded in the text.

4. The claim is **partially supported**, but not fully — i.e., some parts are correct while others are missing or uncertain.

General rule:

Only mark a claim as **True** (supported) if you can clearly point to a sentence or set of sentences in the other article that fully confirm it, leaving **no doubt** about its accuracy. Otherwise, mark it as **False** (not supported).

Figure 10: Instructions for the human metric judgments (ICJ)

bration for support judgments. To best evaluate against these modalities, models should be trained or modified for calibration. In Appendix H, we provided one example for this calibration, however, future work should continue to improve upon this baseline.

I.1 Example Data Applications

Some example data for evaluation are text (Rush et al., 2015; Nallapati et al., 2016; Narayan et al., 2018; Xu and Lapata, 2021; Shapira et al., 2022; Gantt et al., 2024; Walden et al., 2025a), image (Liu et al., 2023b), and video (Krishna et al., 2017; Lin et al., 2024; Hua et al., 2024; Lin et al., 2025; Sanders et al., 2024).

J Extended Related Work

RAG Benchmarks There are several works that tackle the task of video-based RAG (VideoRAG; Jeong et al., 2025; Martin et al., 2025; Ren et al., 2025). However, we focus our evaluation on WikiVideo (Martin et al., 2025) because, unlike other work, the task requires generating long-form text, from a well defined information need, and across multiple videos, which makes properly citing sources a task requirement. This best mirrors the common setting of TextRAG with targeted information seekign queries, multiple information sources, and citations to the information used in generation (Barham et al., 2023; Gupta et al., 2024; Han et al., 2024; Lawrie et al., 2024).

In evaluating their article generation task, Martin et al. (2025) use ROUGE (Lin, 2004), BERTScore (Zhang et al., 2019), AlignScore (Zha et al., 2023), as well as an argument F_1-like metric (Du et al., 2021; Chen et al., 2023; Vashishtha et al., 2024; Gantt et al., 2024) to evaluate answer span alignment. This evaluation fails on three main aspects: the evaluation measures information only at surface level similarity and n-gram overlap, the evaluation did not consider citations, and the evaluation did not check information against the used videos.

Claim Decomposition Claim decomposition methods are a core component of factuality assessments (e.g. Min et al., 2023; Song et al., 2024). Subclaims, that are decomposed from sentences, are argued the appropriate unit of assessment for evidential support because subclaims are easier and less ambiguous to evaluate than sentences. This allows for a finer granularity in the evaluation of individual, automatic facts, instead of evaluating multiple propositions in a single sentence. There are multiple viewpoints on the level of granularity of subclaims (e.g. Gunjal and Durrett, 2024; Wanner et al., 2024a,b), but we follow the level of subclaim granularity in Martin et al. (2025) in order to utilize their subclaim grounding annotations to evaluate the performance of VLMs in claim verification and enable future work toward calibration.

Related Metrics We provide individual sections and appendices to the RAG metrics we implement, but other metrics for evaluating the quality of generations exist in the literature. For example, FaithScore (Jing et al., 2024) evaluates the factuality of claims in a caption based on their corresponding image. The Pyramind Method (Nenkova and Passonneau, 2004), propose abstracting model summaries into Summary Content Units (SCUs). Automatic implementations of this method represent SCUs as embeddings or QA pairs (Gao et al., 2019; Zhang et al., 2024). Thakur et al. (2025) evaluate their TREC RAG task using LLM judgments on the level of support (full/partial/no) between a sentence and citation.