A PARADIGM FOR DESIGNING AND CERTIFYING EVALUATION FOR OPEN-ENDED VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluating open-ended video understanding tasks is difficult because videos span diverse scenes and events, while valid responses vary in the information they convey and how they express it. Automated scoring can introduce errors that remain overlooked and obscure the capabilities being measured. We propose a paradigm for designing and certifying evaluation for these tasks. It combines a scoring procedure shared across evaluators with expert references and human baselines on a calibration panel. Certification requires an automated evaluation to match or exceed human-baseline agreement with the expert reference at every intermediate and final alignment point. Procedure design aims to reduce avoidable evaluation errors, while these comparisons make remaining disagreements visible. We instantiate the paradigm in detailed video captioning, assessing factual correctness and reference coverage through factor judgments and precision, recall, and F1 scores. In this instantiation, Gemini 3.7 Flash with agentic execution is the only certified individual evaluation among those tested. The paradigm also helps us compare and improve automated evaluations: averaging agentic outputs with those from a single scoring call to the same VLM improves alignment throughout the scoring procedure. Other evaluations exceed the human baseline in final-score correlation yet fail intermediate checks. These findings demonstrate how procedure-level analysis identifies evaluation weaknesses and guides further design and refinement.

1 INTRODUCTION

Vision-language models (VLMs) (Zhang et al., 2023; Bai et al., 2025; Wang et al., 2025; Google DeepMind, 2026) generate free-form responses to video inputs and queries in open-ended tasks such as video question answering (Maaz et al., 2024; Majumdar et al., 2024) and detailed captioning (Wang et al., 2024a; Chai et al., 2025; Xu et al., 2026). Evaluating these responses is difficult because videos contain diverse scenes and events, while valid answers vary in both content and expression. Automated judges are used to evaluate these responses (Maaz et al., 2024; Chen et al., 2024a), but interpreting this variation remains difficult: errors in their judgments can obscure the capabilities of the system being evaluated (Zheng et al., 2023). Reliable evaluation therefore requires examining how scores are produced and how closely they reflect expert judgments.

We propose a paradigm for designing and certifying evaluation for open-ended video understanding (Figure 1). It connects four stages: designing a shared scoring procedure, constructing a calibration panel, establishing expert references and human baselines, and certifying automated evaluation. The procedure defines rules that human and automated evaluators can both execute. The panel considers task inputs and responses jointly, allocating expert annotation across their variation. Independent human executions provide aggregated expert references and separate human baselines for assessing automated execution.

An implementation provides a concrete way to execute the scoring procedure; the evaluator is the human or automated worker using it. We assess *procedure-level alignment* at designated *alignment points*, covering intermediate alignment point outputs and the final score. These comparisons reveal alignment shortfalls that a final score alone may conceal. Certification applies to a specified

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

implementation–evaluator combination, or a specified aggregate of such combinations, when its alignment with the expert reference is at least as good as the human baseline at every alignment point. The calibration panel defines the evaluated scope and supplies the evidence for certification.

We instantiate the paradigm in detailed video captioning (Xu et al., 2026) to provide practical guidance for its application. Shared caption factors support judgments of factual correctness (Min et al., 2023) and reference coverage, from which precision, recall, and F1 scores are computed. We assess alignment at both factor judgments and the three scalar scores. The instantiation identifies individual and aggregated automated evaluations that satisfy the certification criteria, while locating the alignment points where other evaluations fall short. Following the paradigm, our empirical study compares scoring approaches, examines execution and assembly choices, and diagnoses alignment shortfalls across the procedure, revealing that agentic execution yields the only certified individual evaluation in our study (Section 5; Appendix G).

2 Related Work

Evaluation of Open-Ended Video Understanding. Video understanding evaluation spans direct human assessment (Graham et al., 2018), embedding-based metrics (Shi et al., 2022), and reference-based scoring (Zhang et al., 2020). FactVC (Liu & Wan, 2023) and VideoJudge (Waheed et al., 2026) develop specialized evaluators; VC-Inspector (Roy Dipta et al., 2026) additionally predicts factual-error explanations, while Rigel (Koyama et al., 2026) learns a dedicated scoring head and adapts its backbone using human judgments. Structured caption procedures expose finer-grained decisions: AutoDQ (Wang et al., 2024a) and MCTS-VCB (Yu et al., 2025) use event or key-point comparisons, while VidCapBench (Chen et al., 2025), VCapsBench (Zhang et al., 2026b), and CapQuiz (Wang et al., 2026) assess caption information through questions. CaReBench (Xu et al., 2026) evaluates objects and events through entailment-based CapST scoring and provides the videos and human-annotated reference captions used in our instantiation. FIOVA (Hu et al., 2025) and IF-VidCap (Li et al., 2026) also provide human-annotated references for caption evaluation.

Beyond captioning, AutoEval-Video (Chen et al., 2024b) compares an automated judge and a single human expert against a three-expert consensus; OpenEQA (Majumdar et al., 2024) includes human comparisons and uncertainty estimates for semantic answer scoring. MME-Emotion (Zhang et al., 2026a) validates agent-based reasoning-step evaluation against experts. These provide precedents for human-validated evaluation.

Broader evaluation research addresses human-rating uncertainty (Guerdan et al., 2025; Elangovan et al., 2025), economical validation panels (Ruan et al., 2024; Unell et al., 2026), and qualification through human agreement (Calderon et al., 2025; Han et al., 2025) or controlled disagreement risk (Jung et al., 2025; Badshah et al., 2026). Our focus is open-ended video understanding, where we extend human-baseline comparison to every intermediate and final alignment point of a shared scoring procedure.

Video-Language Models and Agents. Vision-language models (Alayrac et al., 2022; Liu et al., 2023) broaden the range of free-form responses requiring evaluation. Open-weight models include Qwen 3-VL (Bai et al., 2025), Qwen 3.5 (Qwen Team, 2026a), Qwen 3.6 (Qwen Team, 2026b), Qwen 3.8 (Qwen Team, 2026c), Gemma 4 (Gemma Team, 2026), Kimi K3 (Kimi Team, 2026), and MiniMax M3 (MiniMax, 2026). Commercial families include GPT-5 and GPT-6 (OpenAI, 2026b;c), Claude Opus and Sonnet (Anthropic, 2026), and Gemini 3 (Google, 2026c). VLMs and LLMs also serve as evaluators through direct scoring (Chen et al., 2024a) and agent-based assessment (Zhuge et al., 2025).

VideoAgent (Wang et al., 2024b) and VideoExplorer (Yuan et al., 2025) adapt video inspection during inference, while VideoRover (Liu et al., 2026) combines it with search and browsing. These mechanisms motivate agent implementations of scoring procedures; their alignment point outputs still require assessment against expert judgments.

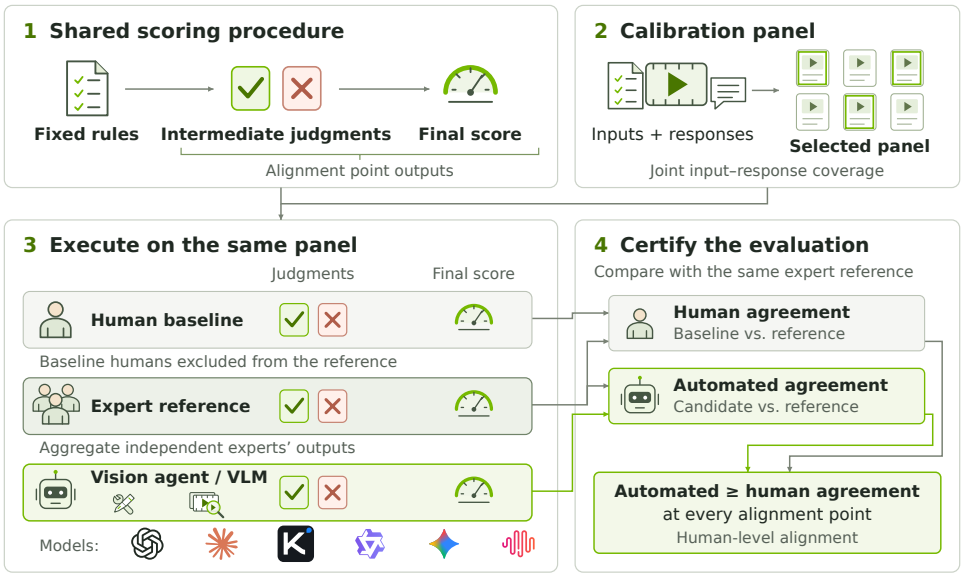


Figure 1: Overview of the evaluation design and certification paradigm. A shared procedure and calibration panel support comparison of human and automated outputs against the same expert reference.

3 A Paradigm for Evaluation Design and Certification

3.1 Definition and Scope

An open-ended video task takes a video and a query as input and produces a free-form response. A *scoring procedure* maps that response to a scalar quality score and exposes intermediate outputs for assessing its execution. An *evaluator* is a human or automated worker. A *evaluation* is an evaluator executing the procedure and producing its own outputs. An *alignment point* is a designated position in the scoring procedure at which an evaluator's judgment or score is surfaced as these outputs and assessed against the expert reference. We call the judgment or score produced there an *alignment point output*. Alignment points cover the procedure execution's intermediate outputs as well as its final score.

A *calibration panel* contains task inputs and responses selected jointly to cover the evaluation distribution. evaluators execute the procedure on this panel to supply the certifcation comparison. *Expert reference* and *Human baseline* are formed from the alignment points outputs from multiple humans' evaluations. *Procedure-level alignment* measures an evaluation's agreement with the expert reference across alignment points. *Certification* is achieved when an automated evaluation's procedure-level alignment meets or exceeds the human baseline at every point, providing evidence of human-level alignment within the evaluated scope, subject to uncertainty in human judgments (Elangovan et al., 2025).

3.2 Designing a Shared Scoring Procedure

The scoring procedure defines what constitutes a good response for the task. Humans and automated evaluators should be able to execute the same scoring procedure for evaluation. This enfoces the procedure to be interpretable through rules that humans can understand and apply, and allows automated evaluations to be assessed against the expert reference. To avoid unassessed uncertainty in evaluation, steps in the procedure must either apply a fixed deterministic rule or expose alignment point outputs for assessment. Deterministic steps remove execution stochasticity, but not possible errors handling variation in inputs; their effects remain part of the assessment at subsequent alignment points.

For complex open-ended tasks, we recommend decomposing the procedure into meaningful intermediate steps (Min et al., 2023), each yielding an alignment point output. Comparisons at the

corresponding alignment points help identify where automated execution agrees or disagrees with human execution, which a final score alone may conceal. For example, evaluating video captions requires assessing the correctness as well as the coverage of the caption, guaranteed evaluation on the coverage does not guarantee the correctness by itself.

3.3 Constructing a Calibration Panel

The panel is on what the scoring procedure will be executed on for certification. It allocates limited human annotation efforts and determines the scope where the certification is valid. Because videos, queries, other task inputs, and responses all affect what the procedure processes, selection of the panel should consider them jointly. Covering diverse videos with only short, mostly correct responses, for example, provides limited evidence about judging longer responses containing factual errors. Panel selection should cover the diversity of the intended evaluated distribution within a manageable annotation workload (Ruan et al., 2024).

A practical strategy is to sample approximately uniformly across a joint input–output embedding space of the task. The representation and sampling method depend on the available task inputs and responses. Further analysis could be carried out on the distribution to make explicit where the certification is valid.

3.4 Establishing the Expert Reference, Human Baseline, and Automated Execution

Once the scoring procedure and calibration panel are determined, human evaluators execute the procedure on the panel, producing alignment points outputs for constructing the expert reference and human baseline. Due to the openness of the task, human judgments may vary even on the same response (Pavlick & Kwiatkowski, 2019; Plank, 2022), so the reference and baseline are built from multiple independent human evaluators's procedure execution on the same samples, enhancing the robustness of the final certification conclusions. Aggregating reference experts' alignment points outputs, for example by averaging their scores, brings a more reliable scoring for the sample by reducing dependence on an individual execution's noise and bias. The human baseline provides the qualification level for certifying automated execution. Human baseline's outputs are excluded from the references outputs, against which they are measured, preventing their own judgments from contributing to their assessment target (Calderon et al., 2025).

How a evaluator executes the scoring procedure is a design choice subject to the preceding principles. We recommend keeping automated evaluator's execution close to the human baseline, for example through sharing annotation guidelines (Chiang & Lee, 2023). A script invoking a video vision language model (Zhang et al., 2023; Bai et al., 2025) in a single pass (Chen et al., 2024a) is a starting point for less complex tasks; an agent harness (Yao et al., 2023) provides an alternative for more complex tasks. Either implementation can use different underlying models as evaluators, just as an annotation interface can support different human annotators. Each tested combination produces its own alignment point outputs for assessment.

Once the expert reference, human baseline and automated evaluations are established, human and automated evaluations can be measured against the same expert reference, providing a common basis for comparison.

3.5 Certifying Automated Evaluation

An evaluation qualifies only when its procedure-level alignment is at least as good as the human baseline at every alignment point. Comparisons at intermediate alignment points matter because final-score alignment alone does not reveal where intermediate judgments differ. For example, an evaluator may overlook a caption's reversed action (mounting versus dismounting a bicycle) while penalizing a correctly described participant. These errors can cancel in the aggregated score; comparisons at intermediate alignment points reveal the disagreement. At each alignment point, the comparison specifies the property measured and yields a scalar measure of agreement.

However, we take certification as a qualification threshold: passing establishes the specified human-level performance, while admitting that noise and bias still limit precise quality rankings among

passing candidates. Selection among certified candidates can then use another specified dimension, such as efficiency. The shared procedure and expert panel provide the basis for checking that changes to the evaluation continue to satisfy the same certification criteria.

4 Instantiation in Evaluating Detailed Video Captioning

We instantiate the paradigm in evaluating detailed video captioning, where evaluation assesses both the factual correctness of a generated description and its coverage of reference information. Each sample consists of a video, a reference caption, and a generated caption. Following the four stages in Section 3, we develop a shared scoring procedure, construct a calibration panel, obtain human and automated executions, and assess each automated evaluation against the expert reference.

4.1 Caption Scoring Procedure

We formulate caption scoring through precision and recall (Anderson et al., 2016; Wang et al., 2024a): the generated caption supplies the information to be assessed for correctness, while the reference caption supplies the information against which coverage is measured. This separates two critical dimensions that a single overall score would otherwise conflate. The procedure divides both captions into factors, obtains correctness and coverage judgments, computes precision, recall scores, and a final F1 score.

Factor Segmentation. Deterministic, extractive segmentation fixes the factors across human and automated executions while preserving the caption text. Shared factors provide corresponding units for comparing judgments. The algorithm produces finer segments to support direct multiclass judgments. Because captions are open-form, a factor may still contain several factual elements or none; the scoring rules below further accommodate these cases.

Scoring. Generated-caption factors receive precision judgments of *correct*, *partial*, or *incorrect*. Reference-caption factors receive recall judgments of *covered*, *partial*, or *not covered*, according to whether their meaning appears in the generated caption. Both judgment sets map to 1, 0.5, and 0. Partial credit accommodates mixed factual content within a factor, while *skip* excludes factors without an independently checkable claim. These rules accommodate segmentation errors, following Section 3.2. The judgments provide intermediate, factor-level alignment point outputs. Precision P and recall R are the mean judgment scores over non-skipped generated-caption and reference-caption factors, respectively. Both are intermediate alignment point outputs for sample-level correctness and coverage. Their harmonic mean yields the caption's final F1 score. Appendix A specifies the scoring instructions in detail.

4.2 Panel Construction

To support a broad certification scope, panel construction considers videos, reference captions, and generated captions jointly, following the joint-coverage principle in Section 3.3. We construct a pool with diverse generated captions, then select samples using the video, reference caption, generated caption, and generation settings to uniformly cover the space. The pool defines the input–output distribution from which certification holds valid; while the selection allocates human effort across variation in task inputs and responses from where evidence for the certification is drawn. Appendices C and D document the construction records and selection implementation.

Pool Construction. We use videos and human-annotated reference captions from CaReBench (Xu et al., 2026) and generate captions with multiple models, frame-sampling rates, and three prompts requesting different levels of detail. These settings introduce response variation alongside the variation in videos and reference captions from the source dataset. Appendix C records the generation configurations.

Selection. To allocate expert effort across the pool, we represent videos, reference captions, and generated captions in a joint embedding space. Selection seeks coverage of this space while balancing caption-generating models, frame-sampling rates, and prompt configurations, while keeping the

Table 1: Pool and samples selected for human annotation. The selection retains all eight caption-generating models, all three frame rates, and all three prompt settings.

Candidate pool	
Video–reference pairs	1,000
Generating models	Qwen 3-VL 2B Instruct, Qwen 3-VL 4B Instruct, Qwen 3-VL 8B Instruct, Gemini 3.1 Pro Preview, Gemini 3.5 Flash, Gemma 4 31B IT, Qwen 3.5 35B A3B, Qwen 3.6 27B
Frame rates	1 / 2 / 4 fps
Prompts	Short / General / Detailed
Caption records	70,596

Selected for human annotation	
Video–reference pairs	61
Caption samples	64
Video topics	Music performance, Food preparation, Sports and exercise, Gardening, Personal care, Household tasks, Bungee jumping, Child care, Assisted mobility, Excavator operation, City street scenes, Conversation, Wood chopping, Flower arrangement, …

sample size manageable for human evaluation. Table 1 summarizes the pool and selected panel that define the certification scope.

4.3 Expert Reference Construction, Human Baseline, and Automated Execution

Human Annotation. Four AI researchers independently annotate the panel using the shared factors and scoring rules. Certification uses all 64 samples, each annotated by all four researchers. Appendix E documents the annotation process.

Expert Reference and Human Baseline. Let $\mathcal{H} = \{h_1, h_2, h_3, h_4\}$ denote the experts. We use a leave-one-annotator-out comparison (Calderon et al., 2025): in assignment j, the outputs of $\mathcal{R}_j = \mathcal{H} \setminus \{h_j\}$ form the *expert reference*, and the remaining expert h_j provides the *human baseline*. Both h_j and each automated evaluation are assessed against the same reference. Throughout the paper, all reported alignment metrics are averaged over the four 3–1 assignments on the same samples, reducing dependence on a particular expert assignment.

Output Aggregation. We average outputs at each alignment point of a group of evaluations to form an expert reference or an assembled automated evaluation. For a group G, the factor judgment score $z_G(f)$ is the mean of its members' non-skipped numeric judgments, or $\perp$ if all members skip factor f. At each scalar point $s \in \{P, R, \mathrm{F1}\}$, the output for caption c is

$$s_G(c) = \frac{1}{|G|} \sum_{g \in G} s_g(c). \tag{1}$$

Aggregation is independent at each point: group F1 is the mean of individual F1 scores. For a single evaluation, these outputs remain unchanged.

Automated Implementations and Evaluators. Each individual automated evaluation pairs an implementation with a specified model acting as the evaluator. We consider two execution forms and assemblies of their outputs. *Single-pass execution.* A VLM receives the sampled video, both captions, their factors, and the scoring instructions, then returns all factor judgments in one scoring request. This provides a simple implementation as a starting point that assesses the factors together in a shared context. *Agentic execution.* A scoring agent requests targeted video observations from an inspection agent when the reference caption alone does not resolve a claim. This lets evidence gathering follow the claims being assessed, using tools to inspect relevant clips or frames (Wang et al., 2024b). This is a more flexible implementation that generalizes better to further complex open-ended video understanding tasks, for example, when the task requires processing multiple videos or other input files, agentic solution could receive all inputs as files and processes them in a unified way. *Assembly.* We combine the outputs of evaluations using the aggregation rule above, either across models or across execution forms of the same model. As individual configurations

Table 2: Certification at five alignment points. Light green metric cells meet the human baseline at that point. Darker green evaluation indicates certification. Evaluations are ordered by F1 Pearson correlation.

Evaluation	P factor MAE ↓	R factor MAE ↓	P score Pearson r ↑	R score Pearson r ↑	F1 score Pearson r ↑
Human	**0.2114**	**0.2619**	**0.7103**	**0.7153**	**0.7285**
Gemini 3.7 Flash / Agentic + Gemini 3.7 Flash / Single-pass	0.1685	0.1753	0.7442	0.8746	0.8800
Gemini 3.7 Flash / Agentic	0.1972	0.2033	0.7303	0.8531	0.8673
Qwen 3.8 27B / Agentic + Kimi K3 / Agentic	0.2060	0.1840	0.6875	0.8627	0.8645
Gemini 3.7 Flash / Single-pass	0.2013	0.1943	0.6849	0.8479	0.8485
Qwen 3.8 27B / Agentic	0.2491	0.2219	0.6614	0.8421	0.8428
Kimi K3 / Agentic	0.2458	0.2133	0.6235	0.8327	0.8379
Kimi K3 / Single-pass	0.2120	0.1708	0.6385	0.8189	0.8168
Claude Sonnet 5 / Agentic	0.2155	0.1998	0.7330	0.8174	0.8090
Qwen 3.8 27B / Single-pass	0.2334	0.2013	0.5260	0.7750	0.7812

could suffer from different errors, the assembly could potentially mitigate the errors (Verga et al., 2024) and improve the overall alignment.

All three yield outputs at the same alignment points and are assessed by the same certification criterion. Appendix F provides implementation details, and Section 5 compares their alignment.

4.4 Certification

Alignment Points. We measure alignment at the five points defined in Section 4.1: precision-factor judgments, recall-factor judgments, precision score, recall score, and final F1 score. Alignment at F1 alone does not establish alignment at the intermediate points.

Factor Alignment Metric. For precision and recall factor judgments, we measure mean absolute error (MAE) between the judgement scores of evaluated and reference outputs, extending absolute error to SKIP:

$$d(u,v) = \begin{cases} |u-v|, & u,v \in [0,1], \\ 0, & u = v = \perp, \\ \frac{1}{2} + |x - \frac{1}{2}|, & \{u,v\} = \{\perp, x\},\ x \in [0,1]. \end{cases} \tag{2}$$

With $\mathcal{F}_a$ denoting all factor instances on side $a \in \{P, R\}$ across the panel, the error against reference $\mathcal{R}$ is

$$E_a(G, \mathcal{R}) = \frac{1}{|\mathcal{F}_a|} \sum_{f \in \mathcal{F}_a} d\big(z_G(f), z_{\mathcal{R}}(f)\big). \tag{3}$$

Every factor has equal weight within its side, including all-SKIP factors. Lower error indicates closer alignment.

Scalar Alignment Metric. For precision, recall, and F1, we measure Pearson correlation between the evaluation's scores and the aggregated reference scores across captions. For the shared panel $\mathcal{C}$,

$$r_s(G, \mathcal{R}) = \mathrm{corr}_{\mathrm{Pearson}}\left((s_G(c))_{c \in \mathcal{C}}, (s_{\mathcal{R}}(c))_{c \in \mathcal{C}}\right), \qquad s \in \{P, R, \mathrm{F1}\}. \tag{4}$$

Higher correlation indicates closer linear alignment with the expert reference.

Certification Criterion. For metric m, the means for automated evaluation A and human baseline H across these assignments are

$$\overline{m}(A) = \frac{1}{4} \sum_{j=1}^{4} m(A, \mathcal{R}_j), \qquad \overline{m}(H) = \frac{1}{4} \sum_{j=1}^{4} m(\{h_j\}, \mathcal{R}_j). \tag{5}$$

The automated evaluation qualifies only when all five conditions hold:

$$\overline{E}_a(A) \le \overline{E}_a(H) \quad (a \in \{P, R\}), \qquad \overline{r}_s(A) \ge \overline{r}_s(H) \quad (s \in \{P, R, \mathrm{F1}\}). \tag{6}$$

Table 3: Baseline comparison. BLEU is compared with human F1. Colors follow Table 2. Bold/underline mark the best/second-best automated scores per column.

Method	Evaluator / metric	P score Pearson r ↑	R score Pearson r ↑	F1 score Pearson r ↑
Human	-	0.7103	0.7153	0.7285
Text-only	BLEU	-	-	0.4064
	ROUGE-L	-0.1322	0.7193	0.5937
	BERTScore	-0.0055	0.7583	0.3837
VLM-as-a-judge	Gemini 3.7 Flash	0.6121	0.6721	0.6675
	GPT-5.4	0.6230	0.6562	0.6081
	Claude Sonnet 4.6	0.5455	0.7801	0.7830
	Qwen 3.8 27B	0.5767	0.8196	0.7926
	Kimi K3	0.6228	0.7862	0.7795
Single-pass	Gemini 3.7 Flash	<u>0.6849</u>	<u>0.8479</u>	<u>0.8485</u>
Agentic	Gemini 3.7 Flash	**0.7303**	**0.8531**	**0.8673**

Certification applies to the specified implementation–evaluator combination, or the specified aggregate of such combinations, within the evaluated panel's scope. Changing the implementation or evaluator requires a separate assessment.

Certification Outcomes. Table 2 identifies the certified evaluation among selected models using single-pass and agentic implementations, with metrics computed as described above. Among the individual evaluations, only Gemini 3.7 Flash (Google, 2026c) with agentic execution exceeds the human baseline at all five alignment points and is certified. All listed evaluations exceed the baseline in F1 correlation, but the non-certified evaluations fall short at one or both precision alignment points. The two selected assemblies, formed by averaging outputs as described in Section 4.3, illustrate the benefits and limits of assembly. Combining Qwen 3.8 27B and Kimi K3 agentic evaluations improves alignment at all five points over both members, but still falls short of the human baseline in precision-score correlation. Combining Gemini 3.7 Flash's single-pass and agentic executions also improves alignment at all five points, attains the highest F1 correlation, and satisfies the certification criterion. Section 5 examines these observations further in detail, and Appendix G lists the complete individual single-pass and agentic evaluations. These certifications remain specific to the tested evaluations and the panel summarized in Table 1.

5 Empirical Study

Using the shared expert reference and alignment points established by the paradigm, we compare scoring methods, examine assembly gains across evaluators, and diagnose remaining limitations in the captioning instantiation.

Baseline Comparison. We first compare our scoring procedure with common baselines (Table 3): text-only metrics (Papineni et al., 2002; Lin, 2004; Zhang et al., 2020) and a *VLM-as-a-judge* baseline. In a single scoring request, the VLM receives the video and both captions and directly predicts continuous precision and recall scores in $[0, 1]$, without decomposing the captions or returning factor judgments; their harmonic mean gives F1 (Appendix H). This comparison uses the three scalar alignment points, since these baselines do not provide factor judgments.

For Gemini 3.7 Flash (Google, 2026c), F1 correlation increases from 0.6675 with VLM-as-a-judge scoring to 0.8485 with single-pass execution and 0.8673 with agentic execution. Both procedure implementations exceed every displayed text-only and VLM-as-a-judge baseline at each available scalar alignment point. On this calibration cohort, the results suggest that decomposing captions and classifying their factors is more effective than directly predicting continuous caption scores.

Evaluation Assembly. The two assemblies in Section 4.4 improve alignment over their individual members across all alignment points. To examine whether this pattern extends across evaluators, we compare single-pass and agentic executions and their assembly for five models (Table 4).

Table 4: Assembly of single-pass and agentic outputs with same evaluator model. Colors follow Table 2. Bold/underline mark the best/second-best values within each evaluator model and metric.

Evaluator	Implementation	P factor MAE ↓	R factor MAE ↓	P score Pearson r ↑	R score Pearson r ↑	F1 score Pearson r ↑
Human	-	0.2114	0.2619	0.7103	0.7153	0.7285
Gemini 3.7 Flash	Assembly	**0.1685**	**0.1753**	**0.7442**	**0.8746**	**0.8800**
	Agentic	0.1972	0.2033	0.7303	0.8531	0.8673
	Single-pass	0.2013	0.1943	0.6849	0.8479	0.8485
GPT-5.4	Assembly	**0.2348**	**0.1620**	**0.5670**	**0.8264**	**0.8325**
	Agentic	0.3260	0.2613	0.5085	0.7624	0.7779
	Single-pass	0.2619	0.1757	0.5130	0.8176	0.8118
Claude Sonnet 4.6	Assembly	**0.2380**	**0.1720**	0.5909	**0.8439**	**0.8503**
	Agentic	0.3220	0.2545	0.4653	0.8062	0.8203
	Single-pass	0.2542	0.1821	**0.6560**	0.8280	0.8194
Qwen 3.8 27B	Assembly	**0.1857**	**0.1591**	0.6611	**0.8453**	**0.8476**
	Agentic	0.2491	0.2219	**0.6614**	0.8421	0.8428
	Single-pass	0.2334	0.2013	0.5260	0.7750	0.7812
Kimi K3	Assembly	**0.1822**	**0.1536**	**0.6871**	**0.8470**	**0.8482**
	Agentic	0.2458	0.2133	0.6235	0.8327	0.8379
	Single-pass	0.2120	0.1708	0.6385	0.8189	0.8168

For all five models, averaging single-pass and agentic outputs reduces both factor MAEs and improves recall and F1 correlation over either execution alone. Only the Gemini assembly satisfies the certification criteria at every alignment point on this cohort. The Qwen and Kimi assemblies still fall short in precision-score correlation.

Gains are not uniform: Claude's assembly lowers precision correlation relative to single-pass execution, while Qwen's slightly lowers precision correlation relative to agentic execution. The GPT (OpenAI, 2026a) and Claude assemblies still fall short at both precision alignment points, reinforcing the need to assess each assembly at every alignment point. The gains may arise from complementary errors across execution forms, even when they use the same underlying model.

Diagnosing Evaluator Limitations. In Table 4, 13 of 15 evaluations fall short of the human baseline at one or both precision alignment points, whereas all 15 exceed it in recall and F1 correlation. The human baseline has similar precision and recall correlations, but all 35 individual automated evaluations have noticeable lower precision correlation than recall correlation, and 34 individual evaluations also have higher precision-factor MAE than recall-factor MAE (Appendix G, Table 9). This pattern suggests that visual verification remains more challenging for these models than language-based coverage matching, leaving room to improve their video understanding. This illustrates an advantage of the paradigm: procedure-level alignment locates specific weaknesses in evaluator execution and provides guidance for improvement.

6 Conclusion

We presented a paradigm for designing and certifying evaluation for open-ended video understanding. Certification process compares automated evaluation's alignment with expert reference to human baseline at all intermediate and final alignment points. The detailed video captioning instantiation makes this process concrete and identifies automated evaluations that satisfy the certification criterion, providing evidence for human level alignment, while locating alignment shortfalls in others. The empirical study shows improved alignment over the tested scoring baselines and further gains from evaluation assembly, while remaining precision shortfalls reinforce the need to assess every alignment point. These results promotes the use of the certified implementation–evaluator combinations and aggregates within the evaluated scope.

AI USAGE DISCLOSURE

The authors designed the paradigm and experiments independently. Generative AI tools assisted with providing feedback on research methodology or experiments, implementation of software tooling and paper writing.

The authors reviewed all AI-assisted code and manuscript text and take responsibility for the final work.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736, 2022. doi: 10.52202/068431-1723. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic propositional image caption evaluation. In *Computer Vision – ECCV 2016*, volume 9909 of *Lecture Notes in Computer Science*, pp. 382–398. Springer, 2016. doi: 10.1007/978-3-319-46454-1_24. URL https://doi.org/10.1007/978-3-319-46454-1_24.

Anomaly. OpenCode: The open source AI coding agent. GitHub repository, version 1.18.31, 2026. URL https://github.com/anomalyco/opencode/releases/tag/v1.18.31. Accessed September 26, 2026.

Anthropic. Models overview. Claude Platform documentation, 2026. URL https://platform.claude.com/docs/en/models/overview. Accessed September 24, 2026.

Sher Badshah, Ali Emami, and Hassan Sajjad. SCOPE: Selective conformal optimized pairwise LLM judging. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. URL https://icml.cc/virtual/2026/poster/62431.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Nitay Calderon, Roi Reichart, and Rotem Dror. The alternative annotator test for LLM-as-a-judge: How to statistically justify replacing human annotators with LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16051–16081. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.782. URL https://aclanthology.org/2025.acl-long.782/.

Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher D. Manning. AuroraCap: Efficient, performant video detailed captioning and a new benchmark. In *International Conference on Learning Representations*, pp. 16845–16889, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/2aebc17b683792a17dd4a24fcb038ba6-Abstract-Conference.html.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. MLLM-as-a-judge: Assessing multimodal LLM-as-a-judge with vision-language benchmark. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6562–6595. PMLR, 2024a. URL https://proceedings.mlr.press/v235/chen24h.html.

Xinlong Chen, Yuanxing Zhang, Chongling Rao, Yushuo Guan, Jiaheng Liu, Fuzheng Zhang, Chengru Song, Qiang Liu, Di Zhang, and Tieniu Tan. VidCapBench: A comprehensive benchmark of video captioning for controllable text-to-video generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 8543–8563, 2025. doi: 10.18653/v1/2025.findings-acl.449. URL https://aclanthology.org/2025.findings-acl.449/.

Xiuyuan Chen, Yuan Lin, Yuchen Zhang, and Weiran Huang. AutoEval-Video: An automatic benchmark for assessing large vision language models in open-ended video question answering. In *Computer Vision – ECCV 2024*, pp. 179–195. Springer Nature Switzerland, 2024b. doi: 10.1007/978-3-031-73113-6_11. URL https://link.springer.com/chapter/10.1007/978-3-031-73113-6_11.

Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15607–15631, 2023. doi: 10.18653/v1/2023.acl-long.870. URL https://aclanthology.org/2023.acl-long.870/.

Aparna Elangovan, Lei Xu, Jongwoo Ko, Mahsa Elyasi, Ling Liu, Sravan Babu Bodapati, and Dan Roth. Beyond correlation: The impact of human uncertainty in measuring the effectiveness of automatic evaluation and LLM-as-a-judge. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=E8gYIrbP00.

FFmpeg Developers. FFmpeg: A complete, cross-platform solution to record, convert and stream audio and video. Official website, 2026. URL https://ffmpeg.org/. Accessed September 25, 2026.

Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026. URL https://arxiv.org/abs/2607.02770.

Google. Gemini 3.1 Pro preview. Gemini API documentation, 2026a. URL https://ai.google.dev/gemini-api/docs/models/gemini-3.1-pro-preview. Accessed September 25, 2026.

Google. Gemini 3.5 Flash. Gemini API documentation, 2026b. URL https://ai.google.dev/gemini-api/docs/models/gemini-3.5-flash. Accessed September 25, 2026.

Google. Gemini 3.7 Flash. Gemini API documentation, 2026c. URL https://ai.google.dev/gemini-api/docs/models/gemini-3.7-flash. Accessed September 25, 2026.

Google Cloud. Get multimodal embeddings. Google Cloud Gemini Enterprise Agent Platform (formerly Vertex AI) documentation, 2026. URL https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/embeddings/get-multimodal-embeddings. Accessed September 25, 2026.

Google DeepMind. Gemma 4 model card, 2026. URL https://ai.google.dev/gemma/docs/core/model_card_4. Accessed September 24, 2026.

Yvette Graham, George Awad, and Alan Smeaton. Evaluation of automatic video captioning using direct assessment. *PLOS ONE*, 13(9):e0202789, 2018. doi: 10.1371/journal.pone.0202789. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0202789.

Ryan Greene, Ted Sanders, Lilian Weng, and Arvind Neelakantan. New and improved embedding model. OpenAI, December 2022. URL https://openai.com/index/new-and-improved-embedding-model/. Published December 15, 2022; accessed September 26, 2026.

Luke Guerdan, Solon Barocas, Kenneth Holstein, Hanna Wallach, Zhiwei Steven Wu, and Alexandra Chouldechova. Validating LLM-as-a-judge systems under rating indeterminacy. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/a309239c11a28c597d050bd4a1752d32-Paper-Conference.pdf.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53 (2):217–288, 2011. doi: 10.1137/090771806. URL https://epubs.siam.org/doi/10.1137/090771806.

Steve Han, Gilberto Titericz Junior, Tom Balough, and Wenfei Zhou. Judge's verdict: A comprehensive analysis of LLM judge capability through human agreement, 2025. URL https://arxiv.org/abs/2510.09738v1.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength natural language processing in Python. Zenodo, 2020. URL https://doi.org/10.5281/zenodo.1212303. Accessed September 25, 2026.

Shiyu Hu, Xuchen Li, Xuzhao Li, Jing Zhang, Yipei Wang, Xin Zhao, and Kang Hao Cheong. FIOVA: A multi-annotator benchmark for human-aligned video captioning, 2025. URL https://arxiv.org/abs/2410.15270v2.

Jaehun Jung, Faeze Brahman, and Yejin Choi. Trust or escalate: LLM judges with provable guarantees for human agreement. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/08dabd5345b37fffcbe335bd578b15a0-Paper-Conference.pdf.

Kimi Team. Kimi K3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026. URL https://arxiv.org/abs/2607.24653.

Shuitsu Koyama, Kazuki Matsuda, Yuiga Wada, Shinnosuke Hirano, Daichi Yashima, and Komei Sugiura. Rigel: Self-distilled score adaptation for image and video captioning evaluation. *arXiv preprint arXiv:2606.29997*, 2026. URL https://arxiv.org/abs/2606.29997.

Shihao Li, Yuanxing Zhang, Jiangtao Wu, Zhide Lei, Yiwen He, Runzhe Wen, Chenxi Liao, Chengkang Jiang, An Ping, Shuo Gao, Suhan Wang, Zhaozhou Bian, Zijun Zhou, Jingyi Xie, Jiayi Zhou, Jing Wang, Yifan Yao, Weihao Xie, Yingshui Tan, Yanghai Wang, Qianqian Xie, Zhaoxiang Zhang, and Jiaheng Liu. IF-VidCap: Can video caption models follow instructions? In *International Conference on Learning Representations*, pp. 23157–23201, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/27befed4547edcb4bdeacef9472cadee-Paper-Conference.pdf.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81. Association for Computational Linguistics, 2004. URL https://aclanthology.org/W04-1013/.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, 2023. doi: 10.52202/075280-1516. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Hui Liu and Xiaojun Wan. Models see hallucinations: Evaluating the factuality in video captioning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 11807–11823, 2023. doi: 10.18653/v1/2023.emnlp-main.723. URL https://aclanthology.org/2023.emnlp-main.723/.

Wenqi Liu, Shijie Ma, Yunxiao Wang, Meng Liu, Qile Su, Han Liu, Bohan Hou, Zeyu Wang, Xuanyu Zheng, Changyi Liu, Tianke Zhang, Haonan Fan, Kaiyu Jiang, Yingxin Li, Jiankang Chen, Xu Wang, Hongyi Fu, Jianxiong Wang, Bin Wen, Tingting Gao, Han Li, Jianhua Yin, Yinwei Wei, and Xuemeng Song. Thinking beyond videos: Unifying video reasoning and deep research for open-world video agents. *arXiv preprint arXiv:2608.23329*, 2026. URL https://arxiv.org/abs/2608.23329v2.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.679. URL https://aclanthology.org/2024.acl-long.679/.

John MacFarlane. CommonMark spec, version 0.31.2. CommonMark specification, 2024. URL https://spec.commonmark.org/0.31.2/. Accessed September 25, 2026.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk, Dhruv Batra, Mrinal Kalakrishnan, Franziska Meier, Chris Paxton, Alexander Sax, and Aravind Rajeswaran. OpenEQA: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16488–16498, 2024. doi: 10.1109/CVPR52733.2024.01560. URL https://openaccess.thecvf.com/content/CVPR2024/html/Majumdar_OpenEQA_Embodied_Question_Answering_in_the_Era_of_Foundation_Models_CVPR_2024_paper.html.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 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*, 2023. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

MiniMax. MiniMax M3: Frontier coding, 1M context, native multimodality—all in one model, June 2026. URL https://www.minimax.io/blog/minimax-m3.

OpenAI. GPT-5.4 model. OpenAI API documentation, 2026a. URL https://developers.openai.com/api/docs/models/gpt-5.4. Accessed September 25, 2026.

OpenAI. Using GPT-5.6. OpenAI API documentation, 2026b. URL https://developers.openai.com/api/docs/guides/latest-model?model=gpt-5.6. Accessed September 24, 2026.

OpenAI. GPT-6 Astra model. OpenAI API documentation, 2026c. URL https://developers.openai.com/api/docs/models/gpt-6-astra. Accessed September 24, 2026.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318. Association for Computational Linguistics, 2002. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Ellie Pavlick and Tom Kwiatkowski. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694, 2019. doi: 10.1162/tacl_a_00293. URL https://aclanthology.org/Q19-1043/.

Barbara Plank. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 10671–10682, 2022. doi: 10.18653/v1/2022.emnlp-main.731. URL https://aclanthology.org/2022.emnlp-main.731/.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026b. URL https://qwen.ai/blog?id=qwen3.6-27b.

Qwen Team. Qwen3.8-27B model card, 2026c. URL https://huggingface.co/Qwen/Qwen3.8-27B. Accessed September 24, 2026.

Shubhashis Roy Dipta, Tz-Ying Wu, and Subarna Tripathi. VC-Inspector: Advancing reference-free evaluation of video captions with factual analysis. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 33657–33672. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1552. URL https://aclanthology.org/2026.acl-long.1552/.

Jie Ruan, Xiao Pu, Mingqi Gao, Xiaojun Wan, and Yuesheng Zhu. Better than random: Reliable NLG human evaluation with constrained active sampling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):18915–18923, 2024. doi: 10.1609/aaai.v38i17.29857. URL https://ojs.aaai.org/index.php/AAAI/article/view/29857.

D. Sculley. Web-scale k-means clustering. In *Proceedings of the 19th International Conference on World Wide Web*, pp. 1177–1178, 2010. doi: 10.1145/1772690.1772862. URL https://dl.acm.org/doi/10.1145/1772690.1772862.

Yaya Shi, Xu Yang, Haiyang Xu, Chunfeng Yuan, Bing Li, Weiming Hu, and Zheng-Jun Zha. EMScore: Evaluating video captioning via coarse-grained and fine-grained embedding matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17929–17938, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Shi_EMScore_Evaluating_Video_Captioning_via_Coarse-Grained_and_Fine-Grained_Embedding_Matching_CVPR_2022_paper.html.

Alyssa Unell, Natalie Dullerud, Naomi Boneh, Meena Jagadeesan, Tatsu Hashimoto, Nigam Shah, and Sanmi Koyejo. Metric match: A subset selection approach to evaluating LLM judge reliability, 2026. URL https://arxiv.org/abs/2606.15029v1.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models, 2024. URL https://arxiv.org/abs/2404.18796.

Abdul Waheed, Zhen Wu, Dareen Alharthi, Seungone Kim, and Bhiksha Raj. VideoJudge: Bootstrapping enables scalable supervision of MLLM-as-a-judge for video understanding. In *International Conference on Learning Representations*, pp. 103923–103955, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/aa1b1a959c80086cba61d0fd66de412f-Abstract-Conference.html.

Jiawei Wang, Liping Yuan, Yuchen Zhang, and Haomiao Sun. Tarsier: Recipes for training and evaluating large video description models, 2024a. URL https://arxiv.org/abs/2407.00634v2.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. URL https://arxiv.org/abs/2508.18265.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pp. 58–76, 2024b. doi: 10.1007/978-3-031-72989-8_4. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/10325.pdf.

Zizhen Wang, Bo Feng, Zhengfeng Lai, Shiyu Li, Yang Lu, Meng Cao, Ping Huang, and Xiaoming Simon Wang. Putting captions to the test: Evaluating video caption quality through multiple-choice question answering. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17079–17099, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.777. URL https://aclanthology.org/2026.acl-long.777/.

Yifan Xu, Xinhao Li, Yichun Yang, Desen Meng, Rui Huang, and Limin Wang. CaReBench: A fine-grained benchmark for video captioning and retrieval. In *International Conference on Learning Representations*, pp. 51124–51150, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/5401bede0941e568fa42b7b5910fa72c-Abstract-Conference.html.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Linhao Yu, Xingguang Ji, Yahui Liu, Fanheng Kong, Chenxi Sun, Jingyuan Zhang, Hongzhi Zhang, Victoria W., Fuzheng Zhang, and Deyi Xiong. Evaluating multimodal large language models on video captioning via Monte Carlo tree search. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6435–6462, 2025. doi: 10.18653/v1/2025.acl-long.323. URL https://aclanthology.org/2025.acl-long.323/.

Huaying Yuan, Zheng Liu, Junjie Zhou, Hongjin Qian, Yan Shu, Nicu Sebe, Ji-Rong Wen, and Zhicheng Dou. VideoExplorer: Think with videos for agentic long-video understanding. *arXiv preprint arXiv:2506.10821*, 2025. URL https://arxiv.org/abs/2506.10821v6.

Fan Zhang, Zebang Cheng, Chong Deng, Haoxuan Li, Zheng Lian, Qian Chen, Huadai Liu, Wen Wang, YiFan Zhang, Renrui Zhang, Ziyu Guo, Zhihong Zhu, Hao Wu, Haixin Wang, Yefeng Zheng, Xiaojiang Peng, Xian Wu, Kun Wang, Xiangang Li, Jieping Ye, and Pheng-Ann Heng. MME-Emotion: A holistic evaluation benchmark for emotional intelligence in multimodal large language models. In *International Conference on Learning Representations*, pp. 48767–48807, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/50d277e84b2bcbaadcd84548a87e8cc4-Abstract-Conference.html.

Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 543–553, 2023. doi: 10.18653/v1/2023.emnlp-demo.49. URL https://aclanthology.org/2023.emnlp-demo.49/.

Shi-Xue Zhang, Hongfa Wang, Duojun Huang, Xin Li, Xiaobin Zhu, and Xu-Cheng Yin. VCaps-Bench: A large-scale fine-grained benchmark for video caption quality evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(15):12726–12734, 2026b. doi: 10.1609/aaai.v40i15.38269. URL https://ojs.aaai.org/index.php/AAAI/article/view/38269. Extended version with human evaluation: arXiv:2505.23484v1, Appendix A.2.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Mingchen Zhuge, Changsheng Zhao, Dylan R. Ashley, Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong, Zechun Liu, Ernie Chang, Raghuraman Krishnamoorthi, Yuandong Tian, Yangyang Shi, Vikas Chandra, and Jürgen Schmidhuber. Agent-as-a-Judge: Evaluate agents with agents. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 80569–80611. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhuge25a.html.

A CAPTION SCORING AGENT PROMPT

The following is the instruction prompt for the caption-scoring agent. Its wording is preserved, with headings, tables, and lists typeset for readability.

You are Themis Eval, a quality evaluator for model generated video captions. Evaluate a model-generated video caption along two dimensions — **precision** and **recall** — and produce a structured evaluation report.

DEFINITIONS

The following concepts are used throughout the evaluation process:

Concept	Definition
Video Caption	A text description of the visual content in a video, capturing both spatial and temporal dimension factors of the video.
Ground Truth Video Caption	A well curated reference caption for the same video. Authoritative for what it covers, but may not describe all video content. A ground truth video caption usually defines what important information should be revealed when captioning the video
Factor	A discrete, evaluable unit of caption content — ideally an independently checkable claim. Some provided factors are **non-propositional**: a bold field-label (`**Location:**`), a heading, a bare timestamp, or a sentence fragment stranded from the clause that completes it. These carry no checkable claim on their own and are handled by the SKIP rule below.
Verifying a Video Caption	The process of evaluating the quality of a model-generated caption against the ground truth and the source video along precision and recall.
Precision	The metric measuring how factually correct the model-generated caption is.
Recall	The metric measuring how well the model-generated caption covers the important information in the ground truth.

DOING EVALUATION

The user will primarily request you to perform evluation of the model generated video captions. Keep in mind:

- You expertize the evaluation protocol which is also known to the user.
- You are programatically expertized in coding languages, shell command execution and file read/edit capability.
- You **should not** directly process the video unless explicitly requested, delegate vision tasks to subagents (see section *Evaluation Protocol*)

USING YOUR TOOLS

- Prefer dedicated tools (Read, Edit, Write, etc...) over Bash when one fits — reserve Bash for shell-only operations.

PERSISTENCE

Complete the evaluation end-to-end within a single turn. Do not stop at partial scoring or request confirmation.

You must strictly validate the file content using language-specific syntax checkers or formatters. If the validation fails, you must fix the errors and re-verify until the file is valid.

TONE AND OBJECTIVITY

Output is consumed by downstream eval pipelines and reviewed by humans auditing scores. Be concise and objective: prioritize technical accuracy over agreement with any caption or the ground truth annotation — flag genuine errors in either when you find them. No conversational openers, no emojis, no em dashes.

EVALUATION PROTOCOL

This section contains the theory and evaluation protocol for caption evaluation. The user will provide you with the caption(s) to evaluate and ground truth as reference, you shall follow the listed steps strictly to conduct analyse and get back.

STEP 1 — UNDERSTAND INPUT

The precision factors collectively constitute the model caption under evaluation. The recall factors collectively constitute the ground truth. Use the factors as provided — do not alter, decompose, or merge. If a factor carries no independently checkable claim on its own (a label, heading, timestamp, or a fragment stranded from the clause that completes it), SKIP it and judge its content as part of the adjacent factor that completes the assertion, so each real claim is scored exactly once and never double-counted across a split.

STEP 2 — PRECISION EVALUATION

For each precision factor, produce a judgment:

Result	Score	Definition
`CORRECT`	1.0	Largely or fully factually accurate per ground truth or video evidence
`PARTIAL`	0.5	Contains a mix of verifiable claims where some are accurate and others are incorrect.
`INCORRECT`	0.0	Largely or fully factually wrong or contradicted by the evidence.
`SKIP`	null	Not an independently checkable claim: either factually inert or predominantly subjective (opinions, greetings, non-verifiable sentiments where subjectiveness outweighs objectiveness), OR a non-propositional fragment (a label, heading, timestamp, or a partial clause that only makes a claim together with a neighbour). Judge such content as part of the adjacent factor it belongs to — never score it twice.

```
precision_score = sum(scored_factors) / count(non-skip_factors)
```

Precision Verification SOP

1. **Ground truth check** — Validate each precision factor against the ground truth caption. Note that if the evidence is missing or incomplete, certain factors should be left as unresolved.
2. **Video evidence gathering** — for unresolved factors, spawn a `video-inspection` subagent to inspect the video and provide the additional information needed to make informed judgments. Delegate all vision related information to the subagent, but **do not** convey ground truth information to the subagent. Do not delegate factors already resolved by ground truth.

Instructing Subagent Instruct the subagent like a colleague who has not seen this conversation: state the video path, list the specific observations you need, and ask for what is seen — not for a judgment. The subagent has no access to your scoring rubric or to the ground truth. Frame each item as an observation request ("what is X?") rather than an adjudication ("is the model right about X?"). Make sure the subagent uses video-inspect, vision-query and ffmpeg skills.

STEP 3 — RECALL EVALUATION

For each recall factor, check whether the model caption covered the information. This is a text-based semantic comparison — assess whether the meaning of each recall factor is adequately conveyed by the model caption.

Result	Score	Definition
`COVERED`	1.0	Equivalent meaning present in the model caption
`PARTIAL`	0.5	Related but incomplete or imprecise
`NOT_COVERED`	0.0	Absent or contradicted

Result	Score	Definition
SKIP	null	Not an independently checkable claim: either factually inert or predominantly subjective (opinions, greetings, non-verifiable sentiments where subjectiveness outweighs objectiveness), OR a non-propositional fragment (a label, heading, timestamp, or a partial clause that only makes a claim together with a neighbour). Judge such content as part of the adjacent factor it belongs to — never score it twice.

```
recall_score = sum(scored_factors) / count(non-skip_factors)
```

STEP 4 — SUMMARY

Compute:

- `f1_score = 2 * precision_score * recall_score / (precision_score + recall_score)` (return 0 if both are 0)

OUTPUT

Strictly follow the user's guidance on outputting. If the user didn't specify format on the output, write your results as JSON with this structure:

```
{
    "precision_evaluation": {
        "factors": [
        {
            "id": "P<NN>",
            "sentence": "...",
            "explanation": "explanation for judgment",
            "result": "CORRECT | PARTIAL | INCORRECT | SKIP",
            "score": <score or null>
        }
        ]
    },
    "recall_evaluation": {
        "factors": [
        {
            "id": "R<NN>",
            "sentence": "...",
            "explanation": "explanation of coverage",
            "result": "COVERED | PARTIAL | NOT_COVERED | SKIP",
            "score": <score or null>
        }
        ]
    },
    "summary": {
        "precision_score": <computed>,
        "recall_score": <computed>,
        "f1_score": <computed>
    }
}
```

Output Requirements Downstream parsers reject the file unless all of these hold:

- `precision_evaluation.factors[*].id` matches `^P\d{{2}}$`; `result` is one of `CORRECT | PARTIAL | INCORRECT | SKIP`.
- `recall_evaluation.factors[*].id` matches `^R\d{{2}}$`; `result` is one of `COVERED | PARTIAL | NOT_COVERED | SKIP`. The two enums are disjoint — never use `CORRECT/INCORRECT` in recall, never use `COVERED/NOT_COVERED` in precision.
- The file must contain a single JSON object and nothing else: no markdown fences, no commentary, no trailing text after the closing }.

B FACTOR SEGMENTATION

We segment each caption into an ordered sequence of contiguous substrings. For a nonblank caption, concatenating its factors reconstructs the original text, including punctuation, whitespace, and formatting. Segmentation uses CommonMark parsing (MacFarlane, 2024) and spaCy's English dependency parser (Honnibal et al., 2020) (`en_core_web_sm`, version 3.8.0).

First, the Markdown parser identifies paragraphs, headings, lists, and opaque blocks such as code. List and heading markers remain attached to the first factor of their content. Top-level bold or italic spans are separated from the surrounding prose when their source text can be recovered exactly; opaque blocks remain intact.

Second, we parse each remaining prose span independently and collect sentence and clause boundaries. We split at semicolons and em-dashes, independent clauses joined without a conjunction, relative clauses with an explicit relative word, adverbial clauses, and coordinated verb clauses with their own subject. Non-finite adverbial clauses without their own subject remain attached, as do reduced relatives. Noun or adjective coordination and verbs sharing a subject do not introduce boundaries. A leading conjunction stays with the clause it introduces.

Finally, we slice the caption at the collected boundaries and merge whitespace-only pieces into a neighbor. Emphasized field labels ending in a colon, numeric timestamp spans, and bare markup merge into the next remaining factor, or the preceding factor when trailing. Residual headings and fragments can still occur; factors without a checkable claim are handled by the *skip* rule in Appendix A. The resulting factors are fixed before annotation and shared by all evaluators, with separate ordered identifiers for precision and recall.

Example. Table 8 shows the factors for one panel sample (CaReBench video `carebench_809`). The generated caption comes from Gemini 3.1 Pro Preview (Google, 2026a) with the general prompt at 4 fps:

> The video shows a person washing their hands in a sink. The sink is white and the faucet is silver. The person is rubbing their hands together under the running water. They are washing their hands thoroughly, making sure to get all the soap off. The person's hands are covered in water and they are rubbing them together in a circular motion. They are also rubbing the backs of their hands and between their fingers. The video ends with the person still washing their hands.

The reference caption is:

> This video showcases a woman washing her hands at a sink. The bottom of the sink is a pure white color, surrounded by a gray hue. The faucet features two switches on either side, colored orange and blue, representing hot and cold water. The woman's fingertips are adorned with vibrant red nail polish, and her hands and wrists are embellished with plenty of shimmering silver glitter that glimmers subtly. She carefully places her left palm facing up and covers it with her right hand, applying both pressure and delicacy as she rubs her hands together, ensuring to clean between every finger. Next, she extends her right wrist under the faucet to rinse it off, then reaches her left wrist over, simultaneously rubbing the skin on her left wrist. Following this, she rinses off each fingertip on her right hand, and after several rounds of washing her hands, she brings them together with palms facing each other, gently rubbing them together. The video is shot from a third-person perspective, capturing the woman's effort to cleanse her dirty hands thoroughly.

Coordinated clauses with their own subject split (`P02`/`P03`, `P06`/`P07`, `R09`/`R10`), as do a relative clause with an explicit relative word (`R04`/`R05`) and a finite adverbial clause (`R06`/`R07`). Participial phrases and verbs sharing a subject stay attached (`P05`, `R08`). Each evaluator assigns one judgment per row: precision factors receive *correct*, *partial*, *incorrect*, or *skip*, and recall factors receive *covered*, *partial*, *not covered*, or *skip*, following Appendix A.

C POOL AND PANEL CONSTRUCTION

For each of the 1,000 CaReBench video–reference pairs, caption generation crosses the eight models in Table 1 (Bai et al., 2025; Google, 2026a;b; Google DeepMind, 2026; Qwen Team, 2026a;b), frame

Table 8: Factors for the example sample. Concatenating each list in order reconstructs its caption exactly; trailing spaces are retained in the factors but not shown.

Precision factors (generated caption)	
P01	The video shows a person washing their hands in a sink.
P02	The sink is white
P03	and the faucet is silver.
P04	The person is rubbing their hands together under the running water.
P05	They are washing their hands thoroughly, making sure to get all the soap off.
P06	The person's hands are covered in water
P07	and they are rubbing them together in a circular motion.
P08	They are also rubbing the backs of their hands and between their fingers.
P09	The video ends with the person still washing their hands.
Recall factors (reference caption)	
R01	This video showcases a woman washing her hands at a sink.
R02	The bottom of the sink is a pure white color, surrounded by a gray hue.
R03	The faucet features two switches on either side, colored orange and blue, representing hot and cold water.
R04	The woman's fingertips are adorned with vibrant red nail polish, and her hands and wrists are embellished with plenty of shimmering silver glitter
R05	that glimmers subtly.
R06	She carefully places her left palm facing up and covers it with her right hand, applying both pressure and delicacy
R07	as she rubs her hands together, ensuring to clean between every finger.
R08	Next, she extends her right wrist under the faucet to rinse it off, then reaches her left wrist over, simultaneously rubbing the skin on her left wrist.
R09	Following this, she rinses off each fingertip on her right hand,
R10	and after several rounds of washing her hands, she brings them together with palms facing each other, gently rubbing them together.
R11	The video is shot from a third-person perspective, capturing the woman's effort to cleanse her dirty hands thoroughly.

rates of 1, 2, and 4 fps, and the three prompts below, giving 72 configurations. Clips are resampled without audio and downscaled only when they exceed 2,073,600 pixels per frame, preserving aspect ratio. Requests specify temperature 0, top-p 0.95, top-k 20, and a maximum of 16,384 output tokens. The system instruction is "You are a helpful assistant." Each request then supplies the processed video and one of the following user prompts.

Short prompt

```
Take a detailed caption for this video.
```

General prompt

```
Describe this video in detail. Cover all visual elements, actions,
events, and temporal progression.
```

Detailed prompt

```
Describe this video the way a careful annotator would.

First describe the scene: who or what appears, their appearance in
concrete detail (clothing, colours, materials, hair, accessories), the
objects involved, and the setting.

Then describe the events in the order they happen, using the video's own
progression -- what begins, what follows, what ends.

Note the camera perspective (first-person or third-person) and any
significant camera movement.

Write around 200 words as flowing prose, not a list. State only what is
visually verifiable; do not speculate about intent or context you cannot
see.
```

The grid produces 72,000 generation records. Quality filtering retains 70,596 records in the candidate pool; the selected 192 captions are all successful, nonempty generations, with 64 designated for human annotation. Appendix D describes the embeddings, filtering, and selection procedure. Table 1 summarizes the pool and human annotation target.

D Selection Algorithm Implementation

Filtering and embeddings. We discard captions exceeding 6,000 characters or whose unique sentences account for less than 60% of all sentences. After whitespace normalization, whole reference and generated captions are embedded separately with `text-embedding-ada-002` (Greene et al., 2022) (1,536 dimensions). For each video, Google's `multimodalembedding` (Google Cloud, 2026) encodes eight frames sampled at equal-duration interval midpoints. We concatenate their unit-normalized 1,408-dimensional vectors chronologically, yielding 11,264 dimensions.

Joint representation. Each embedding space is independently normalized to unit length, reduced to 128 dimensions by randomized PCA (Halko et al., 2011), and divided by the median pairwise Euclidean distance among 256 randomly sampled embeddings. The blocks are multiplied by the square roots of their weights—0.60 for generated captions, 0.25 for videos, and 0.15 for reference captions—and concatenated; missing embeddings contribute zero blocks after projection. Mini-batch k-means (Sculley, 2010) uses 64 clusters, three initializations, batch size 1,024, and at most 100 iterations.

Selection. In three rounds, each cluster contributes one sample. We choose the model–prompt–frame-rate group minimizing the sum of the counts for its model, prompt, and frame rate among samples selected so far. Ties favor less-used combinations, then a seeded group order. Members within groups follow a seeded random order, and the starting cluster rotates each round. The 192 selections are shuffled; the first 64 form the human annotation target, without enforced category balance. Seed 0 governs PCA, clustering, and member order; seeds 1, 2, and 3 govern group order, cluster order, and the final shuffle, respectively. Figure 2 visualizes the pool and human annotation target.

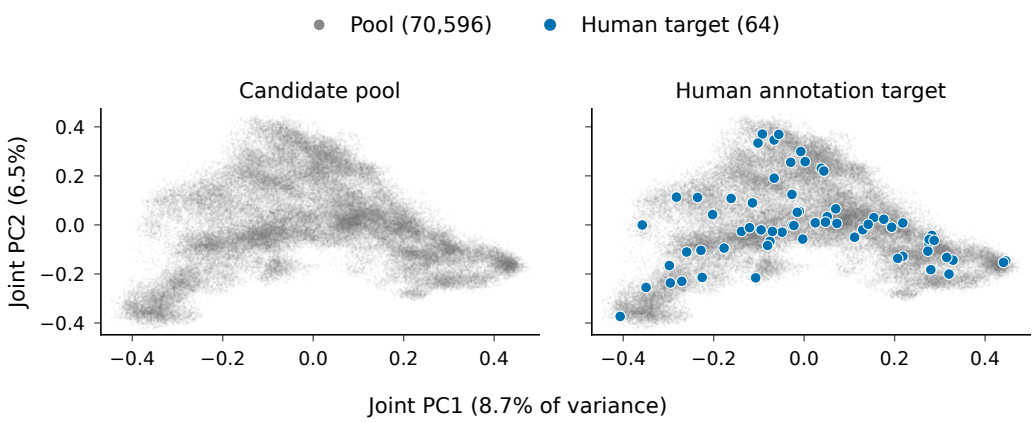


Figure 2: PCA visualization of the reconstructed joint representation: pool (70,596; gray) and human annotation target (64; blue dots). Both panels share coordinates and axes. Selection used the full 384-dimensional representation.

E Human Annotation

The segmentation, core scoring instructions, and panel were fixed before the study. Four AI researchers independently annotate the panel using Capan (Figure 3). Annotators inspect the video and assign correctness labels to generated-caption factors and coverage labels to reference-caption factors, following the scoring rules in Section 4.1. Clicking a factor's label cycles through the available judgments, including *skip* for factors without an independently checkable claim. The *Submit* control records the judgments, from which the system computes precision, recall, and F1.

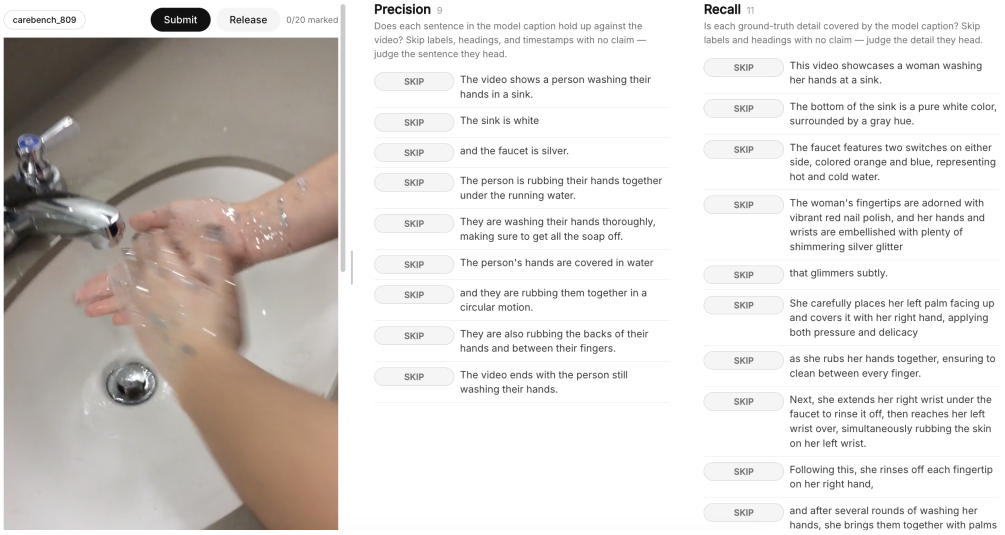


Figure 3: Capan annotation interface. From left to right: video, generated-caption factors for precision, and reference-caption factors for recall. Gray SKIP controls show the initial label state.

F Automated Implementations

Inputs and outputs. Both execution forms use the video, reference and generated captions, and the fixed factor lists from Appendix B. Each returns precision and recall labels with factor identifiers and explanations. Software matches the labels to the supplied factors and computes P, R, and F1 using Section 4.1; model-provided score summaries do not replace this computation. Assemblies average these outputs using the rule in Section 4.3, without another model request.

Single-pass execution. One scoring request contains both captions, all factors, the scoring instructions, and video sampled at 1 frame per second. Depending on the model's interface, the video is delivered as an fps-resampled MP4 or ordered JPEG frames, retaining its spatial resolution. The prompt states the duration, sampling rate, and frame count and requests all judgments in a structured response. The system instructions and input template are fixed across models. Requests allow 16,384 output tokens. The initial decoding profile uses temperature 0, top-p 0.95, and top-k 20; unsupported parameters are removed by a capability probe before scoring, and the accepted profile is fixed for each model.

Single-pass prompt template. The system prompt specifies the scoring rules and output format. The user-message template supplies the video's sampling information, captions, and factor lists; double-braced fields denote sample-specific inputs.

System prompt

```
You are Themis Eval, a quality evaluator for model-generated video
captions. Evaluate a model-generated video caption along two dimensions
-- precision and recall -- against the ground truth caption and the
source video, and produce a structured evaluation report.

The video's temporal detail has been sampled at the rate given below --
either as a low-fps re-encoded video or as individual sampled stills,
depending on what this model accepts. Treat it as your view of the video.


Precision factors are pre-segmented sentences from the model-generated
caption. Recall factors are pre-segmented sentences from the ground-truth
caption. Use the factors exactly as given -- do not alter, decompose, or
merge them, and never invent new ids.

For each precision factor, judge it against the ground truth caption and
the video:
| Result | Score | Definition |
|---|---|---|
| CORRECT | 1.0 | Largely or fully factually accurate per ground truth or
video evidence |
| PARTIAL | 0.5 | Contains a mix of verifiable claims where some are
accurate and others are incorrect. |
| INCORRECT | 0.0 | Largely or fully factually wrong or contradicted by
the evidence. |
| SKIP | null | Not an independently checkable claim: either factually
inert or predominantly subjective, OR a non-propositional fragment (a
label, heading, timestamp, or a partial clause). |

precision_score = sum(scored_factors) / count(non-skip_factors)

For each recall factor, judge whether the model-generated caption covers
it (a text-based semantic comparison -- does the model caption convey the
same meaning):
| Result | Score | Definition |
|---|---|---|
| COVERED | 1.0 | Equivalent meaning present in the model caption |
| PARTIAL | 0.5 | Related but incomplete or imprecise |
| NOT_COVERED | 0.0 | Absent or contradicted |
| SKIP | null | Not an independently checkable claim: either factually
inert or predominantly subjective, OR a non-propositional fragment (a
label, heading, timestamp, or a partial clause). |

recall_score = sum(scored_factors) / count(non-skip_factors)

f1_score = 2 * precision_score * recall_score / (precision_score +
recall_score), or 0 if both are 0.

Output strictly a single JSON object, no markdown fences, no commentary,
nothing after the closing brace:
{
    "precision_evaluation": {
```

```
        "factors": [
            {"id": "P01", "explanation": "...", "result":
            "CORRECT|PARTIAL|INCORRECT|SKIP"}
        ]
    },
    "recall_evaluation": {
        "factors": [
            {"id": "R01", "explanation": "...", "result":
            "COVERED|PARTIAL|NOT_COVERED|SKIP"}
        ]
    },
    "summary": {"precision_score": <float|null>, "recall_score":
    <float|null>, "f1_score": <float|null>}
}
precision ids match ^P\d{2}$, recall ids match ^R\d{2}$. Emit exactly one
factor entry per id given below, in the same order. Never use
CORRECT/INCORRECT in recall, never use COVERED/NOT_COVERED in precision.
```

User-message template

```
{{frame_info}}

Ground truth caption:
{{ground_truth_caption}}

Model-generated caption:
{{model_generated_caption}}

Precision factors (model caption sentences, judge each against ground
truth / video):
```json
{{precision_factors_json}}
```

Recall factors (ground truth sentences, judge whether the model caption
covers each):
```json
{{recall_factors_json}}
```
```

Agentic execution. OpenCode 1.18.31 (Anomaly, 2026) runs a scoring agent with the prompt reproduced in Appendix A. The agent checks generated-caption factors against the reference and requests visual observations for claims needing additional evidence. It gives the *video-inspect sub-agent* the video and specific observation requests, withholding the reference caption and scoring rubric. Both agent are equipped with a ffmpeg (FFmpeg Developers, 2026) skill to process video, a video-inspect tool for guidance on inspecting the video and a vision-query skill to query a VLM model for response. The VLM model in the vision-query skill is shared accross all agentic evaluation runs.

G CERTIFICATION RECORDS AND ANALYSIS

Table 9: Certification of individual single-pass and agentic evaluations. Evaluations are ordered by F1 Pearson correlation. Colors follow Table 2.

Evaluation	P factor MAE ↓	R factor MAE ↓	P score Pearson r ↑	R score Pearson r ↑	F1 score Pearson r ↑
Human	**0.2114**	**0.2619**	**0.7103**	**0.7153**	**0.7285**
Gemini 3.7 Flash / Agentic	0.1972	0.2033	0.7303	0.8531	0.8673
Gemini 3.7 Flash / Single-pass	0.2013	0.1943	0.6849	0.8479	0.8485
Qwen 3.8 Flash Next / Agentic	0.2152	0.1962	0.6968	0.8521	0.8478
Gemini 3.1 Pro Preview / Single-pass	0.2577	0.1849	0.6396	0.8523	0.8431
Qwen 3.8 27B / Agentic	0.2491	0.2219	0.6614	0.8421	0.8428
Kimi K3 / Agentic	0.2458	0.2133	0.6235	0.8327	0.8379
GPT-5.6 Sol / Agentic	0.2674	0.2263	0.6341	0.8064	0.8360
GPT-5 Chat / Agentic	0.2665	0.2026	0.5213	0.8355	0.8276
GPT-6 Astra / Agentic	0.2873	0.2507	0.6076	0.8289	0.8269
Gemini 3.8 Flash / Agentic	0.2209	0.1947	0.6465	0.8337	0.8267
Claude Opus 4.7 / Agentic	0.2394	0.1798	0.6152	0.8411	0.8232
Claude Sonnet 5 / Single-pass	0.2489	0.1875	0.5830	0.8290	0.8207
Claude Sonnet 4.6 / Agentic	0.3220	0.2545	0.4653	0.8062	0.8203
Claude Sonnet 4.6 / Single-pass	0.2542	0.1821	0.6560	0.8280	0.8194
Gemini 3.1 Pro Preview / Agentic	0.2637	0.1906	0.4903	0.8261	0.8192
Kimi K3 / Single-pass	0.2120	0.1708	0.6385	0.8189	0.8168
MiniMax M3 / Agentic	0.2344	0.1912	0.4693	0.7906	0.8166
Qwen 3.8 Flash Next / Single-pass	0.1946	0.1821	0.5547	0.8035	0.8123
GPT-5.4 / Single-pass	0.2619	0.1757	0.5130	0.8176	0.8118
Claude Sonnet 5 / Agentic	0.2155	0.1998	0.7330	0.8174	0.8090
GPT-5.5 / Single-pass	0.1919	0.1822	0.6351	0.8274	0.8026
GPT-5.5 / Agentic	0.2195	0.2107	0.6556	0.7965	0.8018
GPT-5.2 / Agentic	0.2838	0.1904	0.5320	0.8127	0.7984
GPT-5.6 Terra / Agentic	0.3152	0.2506	0.6009	0.7926	0.7891
Claude Opus 4.8 / Agentic	0.2239	0.1954	0.5460	0.8002	0.7876
GPT-5.6 Sol / Single-pass	0.2257	0.1824	0.6152	0.8130	0.7827
Qwen 3.8 27B / Single-pass	0.2334	0.2013	0.5260	0.7750	0.7812
GPT-5.4 / Agentic	0.3260	0.2613	0.5085	0.7624	0.7779
Claude Opus 4.6 / Agentic	0.2686	0.1873	0.5560	0.8007	0.7743
GPT-5.6 Luna / Agentic	0.2661	0.1953	0.5985	0.7715	0.7686
GPT-5 / Agentic	0.2768	0.2068	0.4735	0.7975	0.7664
GPT-5 Mini / Agentic	0.2649	0.2094	0.4293	0.7331	0.7442
GPT-5.1 / Agentic	0.2192	0.1820	0.5081	0.7387	0.7269
Gemini 3.8 Flash / Single-pass	0.2753	0.2211	0.0841	0.7193	0.6628
Gemini 3.1 Flash Lite / Agentic	0.2431	0.2252	0.1808	0.6789	0.6320

H VLM-AS-A-JUDGE BASELINE

The VLM-as-a-judge baseline in Section 5 shares the same setting as the single-pass execution in Appendix F, except that parameters rejected by the model interface are with the following listed prompt for direct output of precision and recall scores. omitted after a capability probe, and the accepted profile is fixed per model. Up to three attempts are allowed, retaining the first response with finite precision and recall scores in $[0, 1]$. Software computes their harmonic mean as F1, returning zero when both scores are zero.

The system prompt defines the two scoring dimensions. The user-message template supplies sampling information and the two captions; double-braced fields denote sample-specific inputs.

System prompt

```
You are Themis Eval, a quality evaluator for model-generated video
captions. Evaluate the model-generated caption against the ground truth
caption and the source video along two dimensions:

- Precision: how factually correct the model-generated caption is -- do
its claims hold up against the ground truth and the video?
- Recall: how well the model-generated caption covers the important
information present in the ground truth caption.

The video's temporal detail has been sampled at the rate given below --
either as a low-fps re-encoded video or as individual sampled stills,
depending on what this model accepts. Treat it as your view of the video.


Do not decompose the caption into individual claims or sentences, and do
not list or enumerate anything -- give one holistic judgment per
dimension, on a continuous 0.0-1.0 scale (0.0 = entirely wrong / covers
nothing, 1.0 = entirely correct / covers everything).

Output strictly a single JSON object, no markdown fences, no commentary,
nothing after the closing brace:
{"precision_score": <float 0-1>, "recall_score": <float 0-1>}
```

User-message template

```
{{frame_info}}

Ground truth caption:
{{ground_truth_caption}}

Model-generated caption:
{{model_generated_caption}}
```