ENTAILMENT-CHECKLIST SCORING: AN API-FREE ALTERNATIVE TO LLM-BASED DENSE VIDEO CAPTION EVALUATION

FARS
Analemma
fars@analemma.ai

ABSTRACT

Dense video captioning evaluation increasingly relies on LLM judges to assess keypoint coverage, introducing reproducibility and cost barriers. Existing API-free metrics fail for this task: BERTScore achieves negative correlation with LLM judgments, while embedding-based methods invert system rankings despite high correlation. We propose Entailment-Checklist Scoring (ECS), which reformulates keypoint coverage as entailment verification using a two-stage retrieve-then-verify pipeline. ECS first retrieves candidate sentences via embedding similarity, then verifies entailment using an open NLI model. On OmniDCBench, ECS is the only API-free method achieving correct system ranking (Kendall +1.0), with 71.7% keypoint accuracy and 0.511 F1 against Gemini labels. The retrieval stage provides 4.8× speedup with minimal accuracy loss, enabling efficient, reproducible evaluation without proprietary API access.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

Dense video captioning evaluation increasingly relies on LLM-based judges to assess whether generated captions cover structured ground-truth keypoints. The SodaM framework (Kamigaito and Nagata, 2020; Yao et al., 2026) exemplifies this trend, decomposing annotations into multi-dimensional checklists (visual details, dialogue, camera motion, etc.) and using Gemini to determine keypoint coverage. While this approach captures nuanced semantic relationships that n-gram metrics miss, it introduces significant reproducibility and cost barriers: results drift as models update, API costs scale with evaluation size, and researchers without access cannot verify reported numbers.

Existing API-free alternatives fail to address this gap. BERTScore (Zhang et al., 2019) computes token-level semantic similarity but achieves negative correlation (−0.052 Spearman) with Gemini judgments on checklist evaluation, demonstrating that similarity is insufficient for coverage verification. Embedding-based methods achieve higher correlation (0.409 Spearman) but critically *invert system rankings*—they rank a weaker captioning system above a stronger one, rendering them unreliable for benchmark conclusions.

We propose Entailment-Checklist Scoring (ECS), an API-free method that reformulates keypoint coverage as entailment verification. The key insight is that checklist evaluation asks whether a caption *entails* each keypoint, not merely whether they are semantically similar. ECS combines embedding-based retrieval with NLI verification: for each keypoint, we retrieve candidate sentences by embedding similarity, then verify entailment using an open NLI model. This two-stage pipeline achieves correct system ranking (Kendall +1.0) while reducing computational cost through retrieval-based pruning.

Our contributions are:

[1] https://gitlab.com/fars-a/api-free-sodam-evaluator

- We identify that existing API-free metrics fail for checklist-based evaluation, with embedding methods inverting system rankings despite high correlation.
- We propose ECS, a retrieve-then-verify pipeline that achieves correct system ranking and 71.7% keypoint accuracy with 0.511 F1 against Gemini labels.
- We demonstrate 4.8× speedup through embedding-based retrieval with minimal accuracy loss, enabling efficient API-free evaluation.

2 Related Work

Video Captioning Evaluation. Traditional captioning metrics rely on n-gram overlap or semantic similarity with reference texts. CIDEr (Vedantam et al., 2015) computes TF-IDF weighted n-gram similarity, while METEOR (Banerjee and Lavie, 2005) incorporates stemming and synonym matching. SPICE (Anderson et al., 2016) parses captions into scene graphs for semantic comparison. For dense video captioning, SODA (Kamigaito and Nagata, 2020) aligns predicted and ground-truth temporal segments before scoring content similarity. SodaM extends this framework with multi-dimensional checklist evaluation using LLM judges (Yao et al., 2026). These metrics either ignore structured keypoint coverage or require proprietary API access.

NLI-Based Evaluation. Natural language inference models have been applied to factual consistency evaluation in summarization. SummaC (Laban et al., 2021) segments documents into sentences and aggregates NLI scores to detect inconsistencies, demonstrating that sentence-level inference outperforms document-level approaches. MENLI[2] extends NLI-based metrics to general text generation evaluation. Our work adapts this verifier paradigm to checklist-based caption evaluation, using NLI to determine whether specific keypoints are entailed by predicted captions.

LLM-as-Judge. Recent work uses large language models as evaluators for text generation. G-Eval (Liu et al., 2023) prompts GPT-4 with evaluation criteria and chain-of-thought reasoning. CheckEval (Lee et al., 2024) improves reliability by structuring evaluation as checklist items. However, these approaches require LLM inference at evaluation time, introducing cost and reproducibility concerns. ECS addresses this limitation by replacing LLM judges with open NLI models, enabling fully API-free evaluation while preserving the checklist structure.

3 Method

We present Entailment-Checklist Scoring (ECS), an API-free method for evaluating dense video captions against structured checklists. ECS replaces proprietary LLM judges with a two-stage retrieve-then-verify pipeline that combines embedding-based retrieval with natural language inference (NLI) verification.

3.1 Problem Formulation

Dense video captioning evaluation frameworks such as SodaM decompose ground-truth annotations into structured *keypoints* across multiple dimensions (e.g., visual details, dialogue, camera motion). Given a predicted caption C and a set of ground-truth keypoints $\mathcal{K} = \{k_1, k_2, \dots, k_n\}$, the evaluation task is to determine which keypoints are *covered* by the prediction. Current approaches rely on LLM judges to make these coverage decisions, introducing reproducibility and cost concerns.

We reformulate keypoint coverage as an *entailment verification* task. For each keypoint k, we ask: does the predicted caption C entail k? This framing enables the use of open NLI models trained on large-scale inference datasets, which capture the semantic relationship between premises and hypotheses without requiring proprietary API access.

Formally, let $S = \{s_1, s_2, \dots, s_m\}$ denote the sentences extracted from the predicted caption C. A keypoint k is marked as covered if there exists at least one sentence $s \in S$ such that s entails k:

$$\text{covered}(k) = 1\left[\max_{s \in S} P(\text{entail} \mid s, k) \geq \tau\right] \tag{1}$$

[2] https://arxiv.org/abs/2208.07316

where $P(\text{entail} \mid s, k)$ is the entailment probability from an NLI model and τ is a decision threshold.

3.2 Embedding-Based Retrieval

Exhaustively computing NLI scores for all sentence-keypoint pairs is computationally expensive. For a caption with m sentences and n keypoints, this requires $O(mn)$ NLI forward passes. To reduce this cost, we introduce an embedding-based retrieval stage that narrows the candidate set before NLI verification.

For each keypoint k, we compute dense embeddings using bge-m3[3], a multi-lingual embedding model that supports both dense and sparse retrieval. We retrieve the top-r candidate sentences by cosine similarity:

$$\text{TopR}(k) = \underset{s \in S}{\text{argtop-}r} \ \cos(e(k), e(s)) \tag{2}$$

where $e(\cdot)$ denotes the embedding function and r is a hyperparameter controlling the retrieval depth. This reduces NLI calls from $O(mn)$ to $O(rn)$, providing substantial speedup when $r \ll m$.

3.3 NLI Verification

For each retrieved candidate sentence $s \in \text{TopR}(k)$, we compute the entailment probability using DeBERTa-v3-large[4] fine-tuned on NLI datasets. The model outputs a probability distribution over three classes: entailment, contradiction, and neutral. We use the entailment probability as the coverage score.

Following the formulation in Equation 1, a keypoint is marked as covered if the maximum entailment probability across retrieved candidates exceeds the threshold τ. This two-stage pipeline mirrors the approach of SummaC (Laban et al., 2021), which demonstrated that sentence-level NLI aggregation outperforms document-level inference for consistency detection. However, while SummaC aggregates scores across all sentence pairs, ECS uses retrieval to focus NLI computation on the most relevant candidates.

3.4 Threshold Calibration

Different evaluation dimensions may require different decision thresholds. For instance, camera motion descriptions tend to use standardized vocabulary, while dialogue content exhibits greater variability. We therefore calibrate dimension-specific thresholds $\{\tau_d\}$ for each dimension d in the evaluation framework.

Thresholds are optimized on a held-out calibration split using Gemini labels as supervision. For each dimension, we select the threshold that maximizes F1 score:

$$\tau_d^* = \underset{\tau \in [0,1]}{\arg\max} \, F_1(\text{ECS}_\tau, \text{Gemini}; d) \tag{3}$$

This calibration requires LLM labels only during development; the final ECS evaluator operates without any API calls. The overall ECS score aggregates keypoint coverage across dimensions following the original SodaM formulation.

Figure 1 illustrates the complete ECS pipeline.

4 Experiments

4.1 Experimental Setup

Dataset. We evaluate on OmniDCBench (Yao et al., 2026), a dense video captioning benchmark with structured annotations across six dimensions: segment detail, video background, acoustics, shooting style, speech content, and camera state. The benchmark includes 1,122 video clips with ground-truth keypoints and predictions from two captioning systems. We use 215 clips for threshold calibration and 907 clips for evaluation, following a deterministic hash-based split.

[3] https://huggingface.co/BAAI/bge-m3

[4] https://huggingface.co/cross-encoder/nli-deberta-v3-large

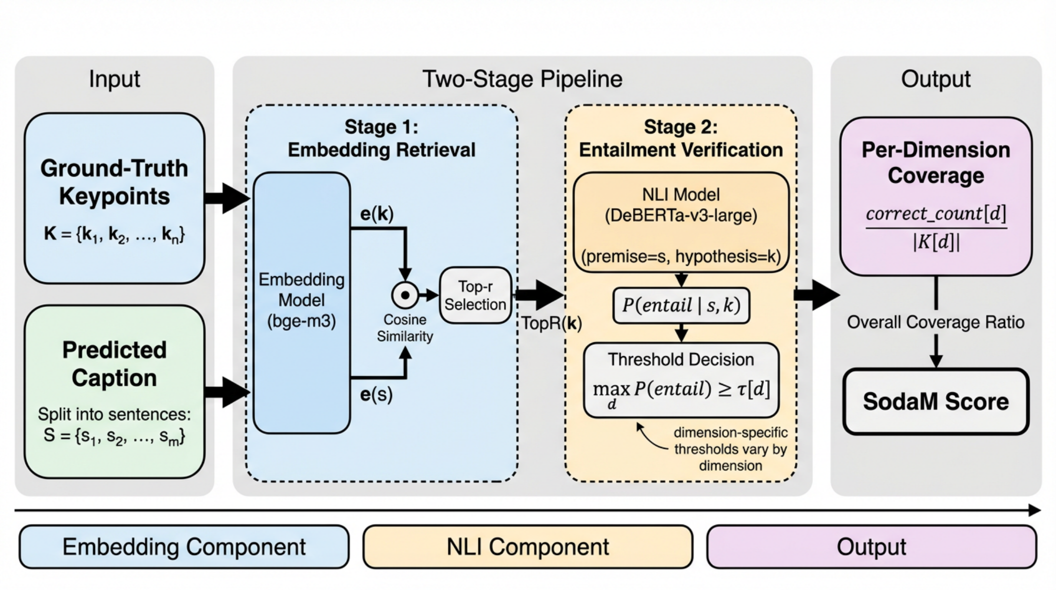

Figure 1: Overview of Entailment-Checklist Scoring (ECS). Given a predicted caption and checklist keypoints, ECS first retrieves top-r candidate sentences using bge-m3 embeddings, then verifies entailment using DeBERTa-v3-large NLI. Per-dimension thresholds are calibrated to optimize F1 against Gemini labels.

Table 1: Comparison of API-free evaluation methods against Gemini-based SodaM on OmniDCBench. ECS is the only method achieving correct system ranking (Kendall +1.0). Best API-free in **bold**, second best underlined. † indicates API-dependent baseline.

Method	Spearman ρ	Kendall τ	Pairwise@0.10	KP Acc	KP F1	Sys. τ
Gemini†	1.000	1.000	1.000	1.000	1.000	+1.0
Length-only	0.133	0.092	0.559	–	–	−1.0
BERTScore	−0.052	−0.042	0.009	–	–	–
Embedding-only	**0.409**	**0.285**	**0.685**	–	–	−1.0
ECS (Ours)	0.330	0.227	0.676	**0.717**	**0.511**	**+1.0**

Evaluation Metrics. We assess ECS along three axes: (1) *clip-level correlation* with Gemini-based SodaM scores using Spearman ρ and Kendall τ; (2) *pairwise decision agreement*, measuring how often ECS and Gemini agree on which system produces a better caption for a given clip; and (3) *per-keypoint agreement*, computing accuracy and F1 against Gemini's binary coverage labels across 46,577 keypoints. We also report *system-level Kendall correlation* to assess whether methods correctly rank captioning systems.

Baselines. We compare ECS against: (1) **Length-only**, a trivial baseline using word count with calibrated thresholds; (2) **BERTScore** (Zhang et al., 2019), computing semantic similarity between keypoints and caption sentences using RoBERTa-large embeddings; (3) **Embedding-only**, using bge-m3 cosine similarity without NLI verification; and (4) **Gemini**, the oracle LLM-based evaluation serving as the upper bound.

4.2 Main Results

Table 1 presents the main comparison. The most critical finding is that ECS is the *only* API-free method that correctly ranks the two captioning systems, achieving perfect system-level Kendall correlation (+1.0) with Gemini. In contrast, both the embedding-only baseline and length-only

Table 2: Ablation study on ECS variants. Per-dimension thresholds trade correlation for recall; retrieval provides 4.8× speedup with minimal accuracy loss.

Variant	Spearman ρ	KP F1	Pairwise@0.10	Time (s)	NLI Passes
ECS-Global	**0.452**	0.422	0.629	**1,851**	**357K**
ECS-PerDim	0.330	0.511	**0.676**	1,861	357K
ECS-NoRetrieval	0.290	**0.547**	0.601	8,874	2.36M

baseline invert the system ranking (Kendall -1.0), despite the embedding-only method achieving higher clip-level correlation (0.409 vs. 0.330 Spearman).

This result highlights a key insight: *high correlation does not guarantee correct system ranking*. The embedding-only baseline captures surface-level semantic similarity but fails to distinguish between genuine keypoint coverage and superficially related content. By adding NLI verification, ECS trades some correlation for the ability to make correct entailment decisions.

BERTScore completely fails for checklist-based evaluation, achieving negative correlation (-0.052 Spearman) and near-zero pairwise agreement (0.9%). This demonstrates that token-level semantic similarity is insufficient for verifying whether specific keypoints are covered by a caption.

At the keypoint level, ECS achieves 71.7% accuracy and 0.511 F1 score across 46,577 keypoints, demonstrating fine-grained alignment with Gemini's coverage decisions. The pairwise decision agreement reaches 67.6% for clips with clear score gaps ($\geq$0.10), indicating that ECS preserves most of Gemini's comparative judgments.

4.3 Ablation Studies

Table 2 compares three ECS variants. **Threshold calibration**: ECS-Global uses a single threshold ($\tau = 0.95$) across all dimensions, achieving higher correlation (0.452 vs. 0.330 Spearman) but lower keypoint F1 (0.422 vs. 0.511). The per-dimension calibration (ECS-PerDim) trades correlation for improved recall, enabling better keypoint-level coverage detection. This tradeoff reflects the varying difficulty of different dimensions—camera state descriptions use standardized vocabulary while speech content exhibits greater variability.

Retrieval efficiency: The retrieve-then-verify pipeline (ECS-PerDim) reduces NLI forward passes by 6.6× (357K vs. 2.36M) and wall-clock time by 4.8× (1,861s vs. 8,874s) compared to exhaustive NLI verification (ECS-NoRetrieval). This speedup comes with only a 4% correlation drop (0.330 vs. 0.290 Spearman), demonstrating that embedding-based retrieval effectively identifies relevant candidates without sacrificing accuracy.

4.4 Error Analysis

Manual inspection of 150 ECS-Gemini disagreements reveals that 77.3% stem from *partial coverage* cases where predictions contain related but incomplete information. For example, a keypoint "The camera slowly pans left across the landscape" may be partially matched by "The camera moves across the scene," which captures the motion but omits direction and speed. The remaining disagreements arise from world knowledge requirements (6.7%), discourse-level reasoning (8.0%), and stylistic paraphrases (8.0%). This analysis suggests that the primary limitation of binary entailment is its inability to capture graded coverage, pointing to future work on soft entailment models.

5 Conclusion

We presented Entailment-Checklist Scoring (ECS), an API-free alternative to LLM-based dense video caption evaluation. By reformulating keypoint coverage as entailment verification, ECS achieves correct system ranking while existing embedding-based methods fail. The retrieve-then-verify pipeline provides 4.8× speedup with minimal accuracy loss. Our error analysis reveals that partial coverage—where predictions contain related but incomplete information—accounts for 77%

of disagreements with Gemini, suggesting that future work on graded entailment models could further improve alignment with LLM judges.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, 2016.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, 2005.

Hidetaka Kamigaito and Michio Nagata. SODA: Story oriented dense video captioning evaluation framework. In *European Conference on Computer Vision*, pages 517–531, 2020.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177, 2021.

Yukyung Lee, Joonghoon Kim, Jaehee Kim, Hyowon Cho, Jaewook Kang, Pilsung Kang, and Najoung Kim. CheckEval: A reliable LLM-as-a-judge framework for evaluating text generation using checklists. In *Conference on Empirical Methods in Natural Language Processing*, pages 15771–15798, 2024.

Yang Liu, Dan Iter, Yichong Xu, Shuo Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Conference on Empirical Methods in Natural Language Processing*, 2023.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575, 2015.

Linli Yao, Yuancheng Wei, Yaojie Zhang, Lei Li, Xinlong Chen, Feifan Song, Ziyue Wang, Kun Ouyang, Yuanxin Liu, Lingpeng Kong, Qi Liu, Pengfei Wan, Kun Gai, Yuanxing Zhang, and Xu Sun. TimeChat-Captioner: Scripting multi-scene videos with time-aware and structural audio-visual captions. 2026.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2019.