Video Comprehension Score (VCS): A Metric for Long-Form Video Description Evaluation

Anonymous submission

Abstract

Existing video description evaluation metrics fail to capture the long-range chronology and semantic alignment essential for long-form descriptions. An effective evaluation metric for long-form descriptions must (i) assess global thematic alignment, (ii) measure local semantic alignment, and (iii) evaluate chronological alignment while detecting corrupted content. We introduce Video Comprehension Score (VCS), a reference-based metric, which directly addresses these evaluation requirements through three components: Global Alignment Score for thematic alignment, Local Alignment Score for local semantic alignment, and Narrative Alignment Score for chronological alignment with adjustable tolerance. We evaluate VCS on two large-scale synthetic datasets designed to test corruption detection and cross-author consistency. VCS consistently outperforms traditional metrics on corruption detection tasks, being the only metric capable of distinguishing valid variations from invalid corruptions. On cross-author consistency tasks, VCS is the only metric that consistently produces scores >80% regardless of which authorial reference is used for evaluation. VCS$_{\text{short}}$, our implementation for short-form descriptions, attains state-of-the-art human correlation on VATEX-EVAL in the 9-ref setting (Kendall's $\tau = 41.5$, Spearman's $\rho = 52.8$) and competitive results in the 1-ref setting (Kendall's $\tau = 30.0$, Spearman's $\rho = 38.1$). These results demonstrate VCS effectiveness for evaluating both long-form and short-form video descriptions.

Introduction

Recent advancements in Large Video Language Models (LVLMs) (Yuan et al. 2025; Shen et al. 2025; Ataallah et al. 2024; Chen et al. 2025) have enhanced automated video comprehension, enabling long-form description generation from long videos. However, exploratory analysis on long videos reveals that LVLMs frequently (i) miss global narrative structure, (ii) fail to capture local events, and (iii) fail to establish temporal connections while hallucinating content. These gaps indicate LVLMs lack true video comprehension, raising the question: how can we evaluate models' video comprehension? Current benchmarks utilize question-answering formats (Wu et al. 2024; Ataallah et al. 2025; Nagrani et al. 2025) or short-form description comparisons for short videos (Wang et al. 2024; Chen and Dolan 2011; Zhou, Xu, and Corso 2018), both inadequate to evaluate models' true video comprehension ability. Evaluating genuine comprehension demands shifting to a methodology that processes long or complex videos and requires models to generate long-form descriptions. This approach renders superficial frame-level analysis insufficient, compelling models to instead generate outputs that capture global narrative structure, detail specific local events, and establish their temporal relationships. The quality of these generated descriptions can then be evaluated against the source video or human-written descriptions.

Description-based evaluation methods fall into four categories, each struggling with fundamental aspects of long-form evaluation. N-gram metrics such as BLEU (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), CIDEr (Vedantam, Zitnick, and Parikh 2015), and ROUGE (Lin 2004) rely on lexical overlap, which not only unfairly penalizes legitimate paraphrases but also rewards superficial word-level matches between descriptions with entirely different global narratives. Embedding-based metrics like BERTScore (Zhang et al. 2020) and SBERT (Reimers and Gurevych 2019) address lexical limitations through semantic similarity but overlook chronological alignment and local finegrained details, allowing misordered events and subtle errors to remain undetected. Multimodal metrics such as EMScore (Shi et al. 2022) enable direct comparison against source video content but struggle with long videos and long-form descriptions. LLM-based metrics such as CLAIR (Chan et al. 2023) and AutoDQ (Wang et al. 2024) provide deeper insights but often do not evaluate the chronology of events while suffering from consistency issues.

To address these limitations, we introduce VCS with three components:

1. **Global Alignment Score (GAS)**: Measures global thematic alignment.
2. **Local Alignment Score (LAS)**: Assesses local semantic alignment.
3. **Narrative Alignment Score (NAS)**: Evaluates chronological alignment using configurable Local Chronology Tolerance (LCT).

VCS combines these three components to provide comprehensive evaluation: GAS and LAS form the Semantic Alignment Score (SAS), which integrates with NAS to produce the final VCS score. We also present VCS$_{\text{short}}$, which applies the same framework to short-form descriptions.

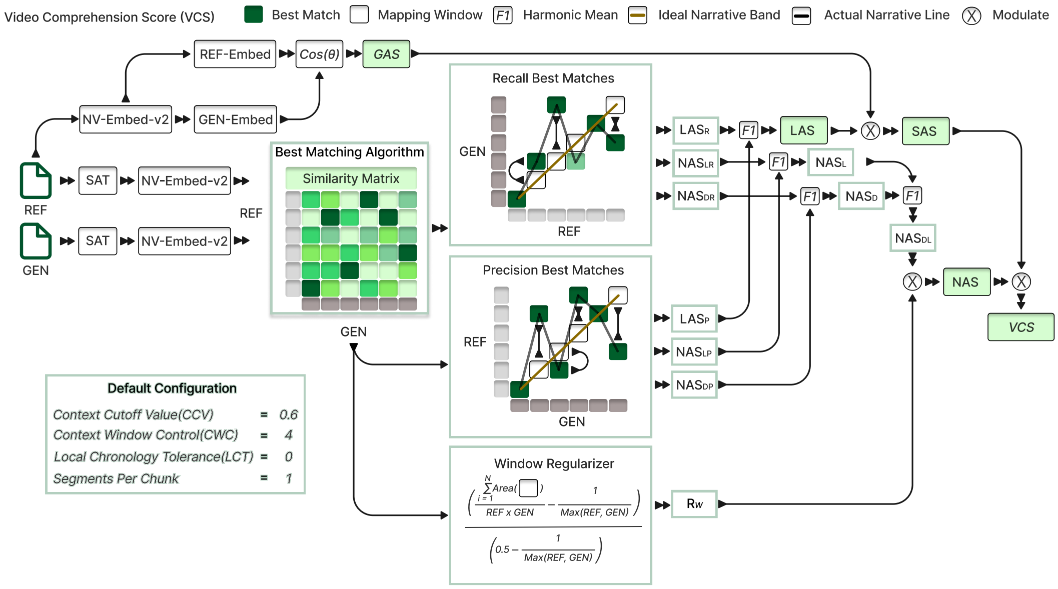


Figure 1: VCS Pipeline. The VCS assesses video descriptions by comparing a reference text (T_{ref}) with a model-generated text (T_{gen}). Both texts are initially segmented by SaT and embedded via NV-Embed-v2. The Global Alignment Score (GAS) is computed from the full text embeddings. For localized analysis, texts are chunked and embedded, forming a similarity matrix. From this, precision and recall-oriented best matches yield the Local Alignment Score (LAS)—the harmonic mean of LAS_P (precision) and LAS_R (recall). The Narrative Alignment Score (NAS) incorporates distance-based (NAS_D) and line-based (NAS_L) assessments. NAS_D and NAS_L are harmonic means of their respective precision and recall components. A Window Regularizer (R_W) refines the NAS. The Semantic Alignment Score (SAS) is derived by modulating GAS with LAS. The final VCS results from modulating the smaller of SAS and the regularized NAS with the larger.

To evaluate VCS, we need datasets of long-form video descriptions paired with quality assessments—but no such datasets exist. Although creating a human-annotated dataset with quality ratings would be ideal, it is impractical: producing and reviewing long-form descriptions is costly and time-consuming, and annotator agreement declines sharply as description length increases. We therefore constructed a large-scale synthetic dataset from the MPII Movie Description Dataset (Rohrbach et al. 2015) via ChatGPT-4o. From this dataset, we derived two test sets: a Corruption Detection Test Set to evaluate VCS ability to distinguish valid variations from invalid corruptions, and a Cross-Author Consistency Test Set to assess robustness across different authorial styles. VCS consistently outperforms traditional metrics on corruption detection tasks. To assess human correlation, we evaluate VCS$_{\text{short}}$ on VATEX-EVAL (Shi et al. 2022), where it achieves state-of-the-art results. These results demonstrate VCS effectiveness for robust evaluation of both long-form and short-form video descriptions.

Our primary contributions include:

- A robust metric (VCS) for both long-form and short-form video descriptions.
- Configurable LCT and chunk size parameters to govern chronological strictness and comparison granularity.

Related Work

Traditional n-gram-based metrics such as BLEU (Papineni et al. 2002), ROUGE (Lin 2004), and METEOR (Banerjee and Lavie 2005) evaluate text generation through lexical overlap and local word order. BLEU measures n-gram precision with brevity penalties, capturing local ordering but missing event-level chronology and struggling with expressive variability. ROUGE variants compute precision, recall, and F1-scores: ROUGE-N evaluates n-gram overlap, ROUGE-L computes Longest Common Subsequence at summary-level, and ROUGE-Lsum calculates sentence-level LCS by splitting at newlines. While preserving sequential information, they miss event-level chronology and remain sensitive to expressive variability. METEOR addresses expressive variability via synonyms and stems, computing recall-weighted F-scores with fragmentation penalties, but similarly misses event-level chronology. CIDEr (Vedantam, Zitnick, and Parikh 2015) uses consensus-based TF-IDF weighting across multiple references but proves impractical for long-form descriptions due to labor-intensive annotation. SPICE (Anderson et al. 2016) evaluates semantic propositions using graph overlaps, effectively handling paraphrasing; however, being designed for static images, it cannot be directly applied to videos and neglects narrative chronology critical for video descriptions.

Embedding-based metrics compare texts in semantic vector spaces, leveraging pretrained models to capture semantic similarity beyond lexical matches. BERTScore computes token-level similarity using contextualized BERT embeddings, aggregating optimal alignments into precision, recall, and F1-scores, while SBERT employs siamese networks with mean-pooling for sentence-level embeddings, but both address expressive variability by recognizing paraphrases yet are constrained by limited context windows, complicating their direct application to long-form descriptions. Recent decoder-based models such as NV-Embed-v2 (Lee et al. 2024), Linq-Embed-Mistral (Choi et al. 2024), SFR-Embedding-Mistral (Meng et al. 2024), and Jasper and Stella (Zhang et al. 2024) leverage autoregressive LLM architectures with bidirectional attention, offering significantly larger context windows and robust global embeddings, excelling at paragraph-level semantic assessments. However, reliance on global embeddings and cosine similarity overlooks local content alignment, detailed information accuracy, and chronological alignment.

Multimodal embedding metrics like EMScore (Shi et al. 2022) and PAC-S (Sarto et al. 2023) employ vision-language models such as CLIP (Radford et al. 2021) to evaluate semantic alignment between visuals and generated captions, addressing expressive variability through direct image-caption comparison, thus bypassing reference captions. EMScore computes video-caption similarity through dual-level matching: coarse-grained alignment between video and caption embeddings, and fine-grained matching between frames and words, calculating precision, recall, and F1-scores via cosine similarity. PAC-S fine-tunes CLIP through positive-augmented contrastive learning with synthetic image-text pairs, then computes cosine similarity scores from the enhanced model. Despite their effectiveness in short-form descriptions, these metrics face computational challenges and methodological limitations when scaling to long-form descriptions.

Recent evaluation approaches increasingly leverage Large Language Models (LLMs), categorized into component-based and holistic judge methods. Component-based methods such as AutoDQ (Wang et al. 2024) extract events from descriptions using ChatGPT, then compute precision and recall through entailment analysis comparing extracted events between texts, while VAD-Score (Dubey and Pack 2025) employs LLMs for semantic extraction and scoring of events, subjects, and objects, effectively addressing expressive variability but missing chronological evaluation. Holistic methods such as CapScore (Li et al. 2024) prompt GPT-4 to score caption similarity and quality, CapArena-Auto Score (Cheng et al. 2025) uses GPT-4o for pairwise evaluation battles, and CLAIR (Chan et al. 2023) employs zero-shot prompting for similarity scores with interpretable reasoning. However, these methods suffer from ambiguity in score calibration, sensitivity to prompting nuances, consistency issues across model versions, limited interpretability, and practical constraints including reproducibility and cost.

Methodology

Figure 1 shows the VCS pipeline, which computes semantic and narrative alignment between reference and generated descriptions to achieve comprehensive long-form video description evaluation.

Global Alignment Score (GAS) VCS computes GAS to capture global thematic alignment between reference and generated descriptions. Given input texts T_{ref} and T_{gen}, we encode each description using NV-Embed-v2. The GAS is computed as:

$$\text{GAS} = \frac{\mathbf{E}_{ref} \cdot \mathbf{E}_{gen}}{\|\mathbf{E}_{ref}\| \|\mathbf{E}_{gen}\|} \tag{1}$$

Text Preprocessing for Chunk-Level Analysis However, GAS lacks sensitivity to finegrained details and chronological consistency. To enable fine-grained analysis, VCS applies punctuation removal and Segment Any Text (SaT) (Frohmann et al. 2024) segmentation to T_{ref} and T_{gen}, producing semantic segments S_{ref} and S_{gen}. We group k consecutive segments into chunks $C_{ref} = \{c_1^{ref}, ..., c_{N_{ref}}^{ref}\}$ and $C_{gen} = \{c_1^{gen}, ..., c_{N_{gen}}^{gen}\}$, embedded using NV-Embed-v2 to yield matrices $\mathbf{E}_{C_{ref}} \in \mathbb{R}^{N_{ref} \times D}$ and $\mathbf{E}_{C_{gen}} \in \mathbb{R}^{N_{gen} \times D}$.

Defining Mapping Windows Before establishing chunk correspondences for fine-grained assessment, VCS constructs a similarity matrix to enable optimal alignments. Given the chunk embedding matrices $\mathbf{E}_{C_{ref}} \in \mathbb{R}^{N_{ref} \times D}$ and $\mathbf{E}_{C_{gen}} \in \mathbb{R}^{N_{gen} \times D}$, we compute similarity matrix $\mathbf{S} \in \mathbb{R}^{N_{ref} \times N_{gen}}$ where $S_{i,j} = \cos(\mathbf{E}_{C_{ref}}[i], \mathbf{E}_{C_{gen}}[j])$ represents the cosine similarity between the i-th reference chunk and j-th generated chunk.

However, before selecting the best match for each chunk, VCS defines Mapping Windows (MW) that constrain the search space within this similarity matrix. This concept stems from empirical observations: when computing pairwise chunk embeddings between identical long-form descriptions, the resulting similarity matrix exhibits a clear diagonal structure where C_1^{ref} maps to C_1^{gen}, C_2^{ref} maps to C_2^{gen}, and so on, with this diagonal pattern stretching or compressing proportionally in brevity or verbity cases. Hence, for given chunk counts N_{ref} and N_{gen}, we define this diagonal search space and call it Mapping Windows—regions where each chunk should ideally map.

We compute slope $s = \max(N_{ref}, N_{gen}) / \min(N_{ref}, N_{gen})$ and base window height $h_{mw} = \lceil s \rceil$. The algorithm creates direct windows by mapping each position i in the shorter sequence to a window of height h_{mw} in the longer sequence, spanning range $[\lfloor i \cdot s \rfloor, \min(\lfloor i \cdot s \rfloor + h_{mw}, N_{longer}))$ with proportional scaling. Reverse windows invert this mapping: for each longer sequence position, we determine which shorter sequence positions include it. We assign Precision Windows (MW_{prec}) and Recall Windows (MW_{rec}) based on sequence direction: when $N_{ref} \geq N_{gen}$, precision uses direct windows and recall uses reverse windows; otherwise, assignments are reversed.

Best Matching Algorithm Having established Mapping Windows within the similarity matrix $\mathbf{S}$, VCS pairs each generated chunk with its best reference counterpart, and vice versa. Naively selecting the highest-similarity reference for each generated chunk leads to semantic collision and semantic ambiguity. Semantic collision occurs when a single chunk exhibits identical similarity to multiple counterparts, while semantic ambiguity arises when a chunk achieves nearly equal similarity with multiple candidates. Both issues stem from repeated events or superficial lexical overlap, obscuring true narrative counterparts. For instance, when repeated events occur, C_1^{gen} may match C_1^{ref} with 0.78 similarity and C_9^{ref} with 0.80 similarity, while C_9^{gen} similarly matches both chunks with high scores. Naive selection pairs $C_1^{gen} \rightarrow C_9^{ref}$ and $C_9^{gen} \rightarrow C_1^{ref}$ based on maximum similarity, when chronologically correct matches should be $C_1^{gen} \rightarrow C_1^{ref}$ and $C_9^{gen} \rightarrow C_9^{ref}$.

VCS resolves these issues through a Best-Matching Algorithm that combines adaptive context filtering with mapping-window constraints. Adaptive context filtering first determines a candidate set around the maximum similarity rather than committing immediately to the top score. Two user-defined parameters control this process: the context cutoff τ_{ctx} (default 0.6) and the context window control k_{ctx} (default 4). Similarities below the cutoff receive no context expansion, while those above it define a context window whose width is computed as

$$w_{ctx} = \frac{(1 - \tau_{ctx}) - (1 - s_{max})}{s_{max} \cdot k_{ctx}}, \tag{2}$$

where s_{max} is the maximum similarity for the chunk. Higher s_{max} values yield broader context windows because small differences between highly similar chunks often reflect positional variations of the same event rather than genuine semantic differences. All candidates within $s_{max} \pm w_{ctx}$ enter the pool for selection. Mapping-window constraints then enforce temporal alignment: among the candidate matches, the algorithm chooses the chunk closest to the mapping-window boundary of the target chunk, breaking ties by similarity. When s_{max} falls below the cutoff, no context expansion occurs, and ties are resolved purely by distance and then similarity.

This combination mitigates both collision and ambiguity. In the above example, both C_1^{ref} (0.78) and C_9^{ref} (0.80) enter C_1^{gen}'s candidate set. Mapping-window distance reveals C_1^{ref} lies in the correct chronological region, and the algorithm selects it over the higher-scoring C_9^{ref}. Similarly, C_9^{gen} pairs with C_9^{ref} rather than C_1^{ref}. Executing this procedure for each generated chunk and each reference chunk yields correspondence sets $M_P = \{(C_j^{gen}, C_{i^*}^{ref})\}$ and $M_R = \{(C_i^{ref}, C_{j^*}^{gen})\}$.

Local Alignment Score (LAS) Having obtained the correspondence sets M_P and M_R from the Best-Matching Algorithm, VCS computes the LAS by averaging cosine similarities of matched chunk pairs. The precision component LAS_P evaluates how well generated text aligns with reference text, while recall component LAS_R evaluates the reverse. LAS is their harmonic mean: $LAS = \frac{2 \cdot LAS_P \cdot LAS_R}{LAS_P + LAS_R}$.

Narrative Alignment Score (NAS) However, LAS remains insensitive to chronological ordering, which inspires us to design NAS that evaluates temporal consistency using the matched pairs M_P and M_R from the Best-Matching Algorithm through distance-based penalties and line-based path analysis.

Distance-based NAS (NAS_D) From the correspondence sets M_P and M_R, NAS_D penalises any match that drifts outside its Mapping Window. In a given orientation let N_{eval} be the length of the timeline being evaluated and N_{opp} the length of the opposite timeline. A local-chronology-tolerance height $h_{LCT} = \lceil N_{eval}/N_{opp} \rceil$ is first derived from the Mapping-Window height, reduced by 1 when $N_{eval} > N_{opp}$ and the fractional part of the ratio lies in $(0, 0.5]$. Each window is then expanded symmetrically by $\tau_{LCT} h_{LCT}$ rows ($\tau_{LCT} \geq 0$ is user-set). For a match at row k with window $[s, e)$ the raw offset is $d(k) = s - k$ when $k < s$ or $k - (e - 1)$ when $k \geq e$; the LCT-adjusted per-chunk penalty becomes $p(k) = \max\{0, d(k) - \tau_{LCT} h_{LCT}\}/N_{eval}$. Summing over all windows yields P_{total}; the worst-case penalty for the same window layout is $P_{max} = N_{eval}^{-1} \sum_j \max(s_j, N_{eval} - e_j)$. The orientation score is $NAS_{D,*} = 1 - P_{total}/P_{max}$, and the two orientations are fused by the harmonic mean

$$NAS_D = \frac{2 \cdot NAS_{D,prec} \cdot NAS_{D,rec}}{NAS_{D,prec} + NAS_{D,rec}}. \tag{3}$$

Line-based NAS (NAS_L) Treating each matched pair as a vertex (x_i, y_i), VCS compares the path that connects them to an ideal narrative band bounded by the shortest and longest feasible paths constrained by the Mapping Windows. Dynamic programming gives the floor length L_{min} and ceil length L_{max}; their vertical increments $\Delta y_{x_i}^{\text{floor}}$ are cached for later clipping. With source length N_{src} and target length N_{tgt} (equal to N_{ref} or N_{gen} depending on orientation) the base window height is $h_{mw} = \lceil N_{tgt}/N_{src} \rceil$. This height in turn defines an LCT kernel ω_0 and its expanded version $\omega_{LCT} = \omega_0 + \kappa \tau_{LCT}$, where the scale κ is h_{mw} or $h_{mw} - 1$ exactly as in the implementation. For consecutive vertices the segment length is

$$\ell_i = \begin{cases} \sqrt{\Delta x_i^2 + \Delta y_i^2}, & 0 \leq |\Delta y_i|^* \leq \omega_0, \\ \sqrt{\Delta x_i^2 + |\Delta y_{x_i}^{\text{floor}}|^2}, & \omega_0 < |\Delta y_i|^* \leq \omega_{LCT}, \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

where $|\Delta y_i|^* = |\Delta y_i|$ if $\tau_{LCT} > 0$ and $|\Delta y_i|^* = \Delta y_i$ when $\tau_{LCT} = 0$. Summing ℓ_i gives the realised length L_{act}. The orientation score is

$$NAS_{L,*} = \begin{cases} 1, & L_{min} \leq L_{act} \leq L_{max}, \\ L_{act}/L_{min}, & L_{act} < L_{min}, \\ L_{max}/L_{act}, & L_{act} > L_{max}. \end{cases} \tag{5}$$

The two orientations are again combined with a harmonic mean to obtain NAS_L.

Window-area Regulariser (R_w) Let W be the Mapping-Window set that spans the longer vertical timeline (MW_{rec} if $N_{gen} > N_{ref}$, otherwise MW_{prec}). With window heights h_j, the covered area is $A_{MW} = \sum_{w_j \in W} h_j$; the full grid area is $A_{timeline} = N_{ref} N_{gen}$. Setting $a =$

$A_{MW}/A_{timeline}$ and $a_{min} = 1/\max(N_{ref}, N_{gen})$, we regularise the window width by

$$R_w = \text{clip}[(a - a_{min})/(0.5 - a_{min}), 0, 1]. \tag{6}$$

Final NAS The distance- and line-based views are first merged: $NAS_{F1} = 2NAS_D NAS_L/(NAS_D + NAS_L)$ (defined as 0 when both inputs are zero). Overly permissive windows are then discounted:

$$NAS_{final} = \begin{cases} (NAS_{F1} - R_w)/(1 - R_w), & NAS_{F1} > R_w, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Thus NAS captures positional fidelity (NAS_D), global path coherence (NAS_L), and penalises inflated scores that could arise from excessively wide Mapping Windows, all while respecting the established notation from the preceding sections.

when $NAS > R_W$ and $R_W < 1$, ensuring meaningful chronological assessment despite length variations.

Final VCS Aggregation The complete VCS integrates semantic and narrative alignment through careful score combination. GAS is modulated by LAS to yield Semantic Alignment Score (SAS):

$$SAS = \frac{GAS - (1 - LAS)}{LAS} \tag{8}$$

when $LAS > 0$ and the numerator is positive, otherwise $SAS = 0$. This ensures high global similarity requires supporting local semantic agreement. The final score synthesizes SAS with NAS_{reg} through adaptive weighting:

The final VCS uses adaptive weighting with intermediate variables:

$$S_{min} = \min(SAS, NAS_{reg}) \tag{9}$$

$$S_{max} = \max(SAS, NAS_{reg}) \tag{10}$$

$$VCS = \frac{S_{min} - (1 - S_{max})}{S_{max}} \tag{11}$$

when both components are positive and the numerator exceeds zero, otherwise $VCS = 0$.

Extension for Short-Form Descriptions (VCS$_{short}$) For shortform description evaluation, VCS$_{short}$ adapts the complete VCS methodology to operate at word-level granularity. Input texts undergo cleaning to remove punctuation and stop words, then tokenization into individual words that serve as fundamental elements. These words replace multi-word chunks in all alignment and scoring processes while maintaining identical metric computation and aggregation logic, enabling consistent evaluation across different description lengths.

Dataset Construction

We construct two synthetic datasets from MPIIMD (Rohrbach et al. 2015) containing 94 movies with 68,000 annotated segments. We extract 1,390 consecutive scene groups (500 words each) and use ChatGPT-4o to synthesize coherent narrative descriptions, yielding our base dataset.

Corruption Detection Test Set

We generate 10 valid variations and 10 invalid corruptions per base description to evaluate VCS ability to distinguish legitimate stylistic changes from content errors. Valid variations include lexical variation, voice transformation, paraphrasing, abstraction (50%/70% reduction), elaboration (130%/150% expansion), action decomposition/aggregation, and attribute injection while preserving narrative integrity. Invalid corruptions introduce content errors through SAO distortion, ID summarization, hallucination (50%/80% unrelated segments), omission (50%/80% deletion), sequence inversion, local/global permutation, and sequence rotation. All transformations employ SAT segmentation while preserving segment content integrity, producing 27,800 test instances.

Cross-Author Consistency Test Set

We use the same 1,390 scene groups with ChatGPT-4o as Author 1 baseline and prompt Grok 3, Claude Sonnet 3.5, and Mistral-Large to generate authorial variations. Each model applies systematic transformations including paraphrasing, voice switching, brevity/verbosity adjustments, action scaling, attribute modification, detail variation, and scene expansion while preserving chronological order and factual content. This produces 5,560 descriptions (1,390 × 4 authors) for evaluating metric robustness across writing styles.

Experiments

We conduct comprehensive experiments to evaluate VCS performance on corruption detection tasks. Our evaluation compares VCS against established video description metrics to assess its ability to distinguish valid narrative variations from invalid corruptions.

Experimental Setup

We evaluate VCS alongside four rule-based metrics: BLEU (Papineni et al. 2002), ROUGE-L (Lin 2004), METEOR (Banerjee and Lavie 2005), and CIDEr (Vedantam, Zitnick, and Parikh 2015), and two embedding-based metrics: BERTScore (Zhang et al. 2020) and SBERT (Reimers and Gurevych 2019). For BERTScore, we use the RoBERTa-base backbone with F1-measure configuration. For SBERT, we employ the all-MiniLM-L6-v2 model for sentence embeddings. VCS uses NV-Embed-v2 for text embeddings with chunk size $k = 3$ and Local Chronology Tolerance $\tau_{LCT} = 0.1$ across all experiments.

The Corruption Detection Test Set provides ground-truth labels where valid variations receive score 1 and invalid corruptions receive score 0. For each metric, we compute scores between base descriptions and their corresponding variations/corruptions, then evaluate classification performance using accuracy, precision, recall, and F1-score with threshold 0.5.

Corruption Detection Results

Table 1 presents the performance of VCS compared to traditional metrics across various text transformations. The transformations are grouped into Valid Test Cases (legitimate variations that should receive high scores) and Invalid Test Cases (corruptions that should receive low scores). For each metric, the top value represents the mean and the bottom value represents the standard deviation.

Human Judgment Correlation Results

Table 2 presents human judgment correlation scores on the VATEX-EVAL dataset, comparing VCS against traditional metrics in both single-reference (1Ref) and multi-reference (9Refs) settings using Kendall's τ_b and Spearman's ρ correlation coefficients.

Cross-Author Consistency Results

Table 3 presents performance comparison of different metrics across various authorial combinations. For each metric, the top value represents the mean and the bottom value represents the standard deviation.

Preparing an Anonymous Submission

This document details the formatting requirements for anonymous submissions. The requirements are the same as for camera ready papers but with a few notable differences:

- Anonymous submissions must not include the author names and affiliations. Write "Anonymous Submission" as the "sole author" and leave the affiliations empty.
- The PDF document's metadata should be cleared with a metadata-cleaning tool before submitting it. This is to prevent leaked information from revealing your identity.
- References must be anonymized whenever the reader can infer that they are to the authors' previous work.
- AAAI's copyright notice should not be included as a footer in the first page.
- Only the PDF version is required at this stage. No source versions will be requested, nor any copyright transfer form.

You can remove the copyright notice and ensure that your names aren't shown by including `submission` option when loading the `aaai2026` package:

```
\documentclass[letterpaper]{article}
\usepackage[submission]{aaai2026}
```

The remainder of this document are the original camera-ready instructions. Any contradiction of the above points ought to be ignored while preparing anonymous submissions.

Camera-Ready Guidelines

Congratulations on having a paper selected for inclusion in an AAAI Press proceedings or technical report! This document details the requirements necessary to get your accepted paper published using PDFLaTeX. If you are using Microsoft Word, instructions are provided in a different document. AAAI Press does not support any other formatting software.

The instructions herein are provided as a general guide for experienced LaTeX users. If you do not know how to use LaTeX, please obtain assistance locally. AAAI cannot provide you with support and the accompanying style files are **not** guaranteed to work. If the results you obtain are not in accordance with the specifications you received, you must correct your source file to achieve the correct result.

These instructions are generic. Consequently, they do not include specific dates, page charges, and so forth. Please consult your specific written conference instructions for details regarding your submission. Please review the entire document for specific instructions that might apply to your particular situation. All authors must comply with the following:

- You must use the 2026 AAAI Press LaTeX style file and the aaai2026.bst bibliography style files, which are located in the 2026 AAAI Author Kit (aaai2026.sty, aaai2026.bst).
- You must complete, sign, and return by the deadline the AAAI copyright form (unless directed by AAAI Press to use the AAAI Distribution License instead).
- You must read and format your paper source and PDF according to the formatting instructions for authors.
- You must submit your electronic files and abstract using our electronic submission form **on time.**
- You must pay any required page or formatting charges to AAAI Press so that they are received by the deadline.
- You must check your paper before submitting it, ensuring that it compiles without error, and complies with the guidelines found in the AAAI Author Kit.

Copyright

All papers submitted for publication by AAAI Press must be accompanied by a valid signed copyright form. They must also contain the AAAI copyright notice at the bottom of the first page of the paper. There are no exceptions to these requirements. If you fail to provide us with a signed copyright form or disable the copyright notice, we will be unable to publish your paper. There are **no exceptions** to this policy. You will find a PDF version of the AAAI copyright form in the AAAI AuthorKit. Please see the specific instructions for your conference for submission details.

Formatting Requirements in Brief

We need source and PDF files that can be used in a variety of ways and can be output on a variety of devices. The design and appearance of the paper is strictly governed by the aaai style file (aaai2026.sty). **You must not make any changes to the aaai style file, nor use any commands, packages, style files, or macros within your own paper that alter that design, including, but not limited to spacing, floats, margins, fonts, font size, and appearance.** AAAI imposes requirements on your source and PDF files that must be followed. Most of these requirements are based on our efforts to standardize conference manuscript properties and layout. All papers submitted to AAAI for publication will be recompiled for standardization purposes. Consequently, every paper submission must comply with the following requirements:

Metric	Valid Test Cases										Invalid Test Cases									
	Lexical Variation	Voice Transformation	Paraphrasing	Low-Abstraction	High-Abstraction	Low-Elaboration	High-Elaboration	Action Decomposition	Action Aggregation	Attribute Injection	Minor Hallucination	Major Hallucination	Minor Omission	Major Omission	SAO Distortion	Summarization	Local Permutation	Global Permutation	Sequence Inversion	Sequence Rotation
BLEU-1	83.1 ±8.5	79.8 ±5.7	69.3 ±5.6	38.3 ±14	23.4 ±11	66.5 ±8.9	55.7 ±7.2	57.2 ±8.9	34.0 ±12	69.5 ±6.5	53.9 ±6.9	43.7 ±7.0	38.9 ±7.1	3.7 ±2.8	49.4 ±4.6	8.1 ±5.4	99.6 ±0.6	99.7 ±0.5	99.7 ±0.5	99.7 ±0.5
BLEU-4	63.2 ±15	52.8 ±10	34.9 ±8.3	18.6 ±11	9.40 ±7.0	36.5 ±14	25.5 ±9.1	25.1 ±10	13.1 ±8.3	62.5 ±8.3	53.5 ±6.9	43.3 ±7.0	38.4 ±7.1	3.60 ±2.7	10.5 ±4.6	2.20 ±2.1	90.0 ±3.7	90.4 ±3.7	90.2 ±3.6	99.2 ±1.0
METEOR	88.1 ±7.1	87.0 ±5.3	65.5 ±8.1	39.2 ±11	29.5 ±8.2	69.7 ±11	63.2 ±8.6	58.7 ±8.8	33.4 ±9.0	86.3 ±4.4	82.6 ±2.4	79.4 ±2.9	47.1 ±4.6	21.0 ±3.7	50.2 ±6.3	17.7 ±4.5	83.7 ±2.0	64.4 ±3.4	58.9 ±3.2	63.2 ±3.7
ROUGE-1	84.0 ±8.1	89.7 ±3.1	74.9 ±4.3	64.0 ±9.6	54.1 ±9.2	77.3 ±6.8	69.9 ±6.0	69.5 ±7.1	58.1 ±9.3	81.8 ±4.6	69.7 ±5.9	60.4 ±7.0	67.0 ±4.3	36.0 ±5.3	51.4 ±4.4	37.2 ±7.0	98.9 ±0.9	99.0 ±0.9	99.0 ±0.9	99.0 ±0.9
ROUGE-4	50.3 ±18	41.0 ±12	20.4 ±7.5	15.7 ±10	8.90 ±6.2	25.6 ±15	16.4 ±8.9	15.3 ±10	9.50 ±6.7	66.3 ±9.3	67.3 ±6.1	58.4 ±7.0	64.2 ±4.8	33.4 ±5.5	2.50 ±2.0	2.90 ±2.3	78.9 ±6.6	79.6 ±6.6	79.2 ±6.6	96.2 ±2.7
ROUGE-L	83.5 ±8.4	75.4 ±6.5	67.0 ±6.0	55.6 ±11	45.6 ±9.8	69.7 ±10	60.6 ±8.4	57.6 ±9.6	45.2 ±10	81.6 ±4.7	69.6 ±6.0	60.3 ±7.0	66.9 ±4.3	36.0 ±5.3	47.2 ±5.4	27.4 ±6.4	65.1 ±2.7	36.5 ±4.6	23.1 ±2.3	53.5 ±3.2
ROUGE-Lsum	83.8 ±8.2	84.7 ±4.1	73.4 ±4.5	62.1 ±10	52.1 ±9.4	76.0 ±7.3	68.4 ±6.3	67.5 ±7.4	55.4 ±9.3	81.8 ±4.6	69.7 ±5.9	60.4 ±7.0	67.0 ±4.3	36.0 ±5.3	50.7 ±4.5	35.2 ±6.8	97.7 ±1.7	97.7 ±1.9	96.7 ±2.4	98.8 ±1.1
$\text{VCS}_{C_1\|LCT_0}$	93.7 ±7.1	95.1 ±3.7	95.8 ±2.8	86.0 ±8.8	80.0 ±10	92.2 ±4.6	87.5 ±5.4	88.5 ±5.9	81.0 ±10	86.4 ±4.7	69.1 ±6.8	53.4 ±14	62.2 ±4.9	1.90 ±4.9	5.00 ±10	62.4 ±18	0.70 ±2.3	3.40 ±4.4	0.10 ±0.8	42.9 ±2.8
$\text{VCS}_{C_1\|LCT_1}$	94.1 ±7.1	96.7 ±2.6	96.7 ±1.8	90.5 ±5.8	86.3 ±6.5	93.5 ±3.9	88.7 ±5.0	90.4 ±5.2	86.9 ±6.9	86.8 ±4.5	70.5 ±7.0	54.9 ±14	68.3 ±4.6	6.70 ±10	5.40 ±10	73.5 ±12	76.1 ±6.5	15.2 ±7.6	43.6 ±2.4	43.0 ±2.8

Table 1: Corruption detection performance across various text transformations. Valid Test Cases represent legitimate variations that should receive high scores, while Invalid Test Cases represent corruptions that should receive low scores. For each metric, the top value represents the mean and the bottom value represents the standard deviation. VCS consistently outperforms traditional metrics by maintaining high scores for valid variations while effectively detecting corruptions.

	1Ref		**9Refs**	
	Kendall τ_b	Spearman ρ	Kendall τ_b	Spearman ρ
BLEU-1	12.2	15.9	28.9	37.0
BLEU-4	12.6	16.4	22.4	29.5
ROUGE	12.5	16.3	23.8	30.9
METEOR	16.4	21.5	27.6	35.7
CIDEr	17.3	22.6	27.8	36.1
BERT-S	18.2	23.7	29.3	37.8
BERT-S++	15.2	19.8	24.4	31.7
EMScore	28.6	37.1	36.8	47.2
PAC-S	00.0	00.0	00.0	00.0
RefPAC-S	**31.4**	**40.5**	38.1	48.8
$\text{VCS}_{C_1\|LCT_0}$	00.0	00.0	00.0	00.0
$\text{VCS}_{C_1\|LCT_1}$	00.0	00.0	00.0	00.0

Table 2: Human judgment correlation scores on VATEX-EVAL dataset.

- Your .tex file must compile in PDFLATEX — (you may not include .ps or .eps figure files.)
- All fonts must be embedded in the PDF file — including your figures.
- Modifications to the style file, whether directly or via commands in your document may not ever be made, most especially when made in an effort to avoid extra page charges or make your paper fit in a specific number of pages.
- No type 3 fonts may be used (even in illustrations).
- You may not alter the spacing above and below captions, figures, headings, and subheadings.
- You may not alter the font sizes of text elements, footnotes, heading elements, captions, or title information (for references and mathematics, please see the limited exceptions provided herein).
- You may not alter the line spacing of text.
- Your title must follow Title Case capitalization rules (not sentence case).
- LATEX documents must use the Times or Nimbus font package (you may not use Computer Modern for the text of your paper).
- No LATEX 209 documents may be used or submitted.
- Your source must not require use of fonts for non-Roman alphabets within the text itself. If your paper includes symbols in other languages (such as, but not limited to, Arabic, Chinese, Hebrew, Japanese, Thai, Russian and other Cyrillic languages), you must restrict their use to bit-mapped figures. Fonts that require non-English language support (CID and Identity-H) must be converted to outlines or 300 dpi bitmap or removed from the document (even if they are in a graphics file embedded in the document).
- Two-column format in AAAI style is required for all papers.

Metric	**Author 1**			**Author 2**			**Author 3**			**Author 4**		
	A1-A2	A1-A3	A1-A4	A2-A1	A2-A3	A2-A4	A3-A1	A3-A2	A3-A4	A4-A1	A4-A2	A4-A3
BLEU-1	45.1	46.3	52.4	46.4	39.4	50.8	43.8	33.7	40.2	53.3	50.2	44.7
	±6.7	±4.3	±9.2	±5.9	±5.0	±5.7	±5.8	±6.8	±9.3	±8.2	±6.3	±6.8
BLEU-4	9.30	9.70	15.2	9.60	8.10	12.0	9.20	7.00	9.70	15.4	11.9	10.7
	±3.5	±2.6	±7.0	±3.5	±2.6	±3.9	±2.6	±2.6	±3.7	±6.9	±3.9	±3.5
METEOR	35.2	42.9	41.6	41.1	43.5	42.8	33.5	29.0	33.9	43.1	38.1	44.8
	±5.0	±3.5	±7.6	±3.8	±3.0	±4.4	±3.3	±3.9	±5.4	±7.5	±6.4	±5.5
ROUGE-1	56.0	57.8	62.1	56.0	55.2	60.2	57.8	55.2	59.4	62.1	60.2	59.4
	±4.8	±3.4	±6.6	±4.8	±4.2	±4.3	±3.4	±4.2	±5.0	±6.6	±4.3	±5.0
ROUGE-4	3.00	3.20	6.60	3.00	3.20	4.50	3.20	3.20	4.40	6.60	4.50	4.40
	±1.8	±1.5	±4.9	±1.8	±1.5	±2.3	±1.5	±1.5	±2.1	±4.9	±2.3	±2.1
ROUGE-L	36.6	40.4	43.8	36.6	34.9	38.3	40.4	34.9	38.8	43.8	38.3	38.8
	±4.9	±3.9	±7.9	±4.9	±4.1	±5.0	±3.9	±4.1	±5.2	±7.9	±5.0	±5.2
ROUGE-Lsum	52.6	55.1	59.2	52.6	51.3	56.4	55.3	51.4	56.2	59.3	56.5	56.0
	±4.7	±3.5	±6.7	±4.8	±4.2	±4.3	±3.5	±4.1	±5.0	±6.7	±4.3	±5.1
VCS$_{C_1\|LCT_0}$	76.3	77.9	74.9	76.3	75.2	76.7	77.9	75.2	76.5	74.9	76.7	76.5
	±10	±7.5	±13	±10	±9.6	±11	±7.5	±9.6	±10	±13	±11	±10
VCS$_{C_1\|LCT_1}$	82.5	80.5	81.6	82.5	79.1	83.8	80.5	79.1	81.6	81.6	83.8	81.6
	±7.7	±6.5	±10	±7.7	±7.7	±8.8	±6.5	±7.7	±8.2	±10	±8.8	±8.2

Table 3: Performance comparison of different metrics across various authorial combinations. For each metric, the top value represents the mean and the bottom value represents the standard deviation.

- The paper size for final submission must be US letter without exception.
- The source file must exactly match the PDF.
- The document margins may not be exceeded (no overfull boxes).
- The number of pages and the file size must be as specified for your event.
- No document may be password protected.
- Neither the PDFs nor the source may contain any embedded links or bookmarks (no hyperref or navigator packages).
- Your source and PDF must not have any page numbers, footers, or headers (no pagestyle commands).
- Your PDF must be compatible with Acrobat 5 or higher.
- Your LaTeX source file (excluding references) must consist of a **single** file (use of the "input" command is not allowed.
- Your graphics must be sized appropriately outside of LaTeX (do not use the "clip" or "trim" command) .

If you do not follow these requirements, your paper will be returned to you to correct the deficiencies.

What Files to Submit

You must submit the following items to ensure that your paper is published:

- A fully-compliant PDF file.
- Your LaTeX source file submitted as a **single** .tex file (do not use the "input" command to include sections of your paper — every section must be in the single source file). (The only allowable exception is .bib file, which should be included separately).
- The bibliography (.bib) file(s).
- Your source must compile on our system, which includes only standard LaTeX 2020 TeXLive support files.
- Only the graphics files used in compiling paper.
- The LaTeX-generated files (e.g. .aux, .bbl file, PDF, etc.).

Your LaTeX source will be reviewed and recompiled on our system (if it does not compile, your paper will be returned to you. **Do not submit your source in multiple text files.** Your single LaTeX source file must include all your text, your bibliography (formatted using aaai2026.bst), and any custom macros.

Your files should work without any supporting files (other than the program itself) on any computer with a standard LaTeX distribution.

Do not send files that are not actually used in the paper. Avoid including any files not needed for compiling your paper, including, for example, this instructions file, unused graphics files, style files, additional material sent for the purpose of the paper review, intermediate build files and so forth.

Obsolete style files. The commands for some common packages (such as some used for algorithms), may have changed. Please be certain that you are not compiling your paper using old or obsolete style files.

Final Archive. Place your source files in a single archive which should be compressed using .zip. The final file size may not exceed 10 MB. Name your source file with the last (family) name of the first author, even if that is not you.

Using LaTeX to Format Your Paper

The latest version of the AAAI style file is available on AAAI's website. Download this file and place it in the TeX search path. Placing it in the same directory as the paper should also work. You must download the latest version of the complete AAAI Author Kit so that you will have the latest instruction set and style file.

Document Preamble

In the LaTeX source for your paper, you **must** place the following lines as shown in the example in this subsection. This command set-up is for three authors. Add or subtract author and address lines as necessary, and uncomment the portions that apply to you. In most instances, this is all you need to do to format your paper in the Times font. The helvet package will cause Helvetica to be used for sans serif. These files are part of the PSNFSS2e package, which is freely available from many Internet sites (and is often part of a standard installation).

Leave the setcounter for section number depth commented out and set at 0 unless you want to add section numbers to your paper. If you do add section numbers, you must uncomment this line and change the number to 1 (for section numbers), or 2 (for section and subsection numbers). The style file will not work properly with numbering of subsubsections, so do not use a number higher than 2.

The Following Must Appear in Your Preamble

```
\documentclass[letterpaper]{article}
% DO NOT CHANGE THIS
\usepackage[submission]{aaai2026} % DO NOT CHANGE THIS
\usepackage{times} % DO NOT CHANGE THIS
\usepackage{helvet} % DO NOT CHANGE THIS
\usepackage{courier} % DO NOT CHANGE THIS
\usepackage[hyphens]{url} % DO NOT CHANGE THIS
\usepackage{graphicx} % DO NOT CHANGE THIS
\urlstyle{rm} % DO NOT CHANGE THIS
\def\UrlFont{\rm} % DO NOT CHANGE THIS
\usepackage{graphicx}  % DO NOT CHANGE THIS
\usepackage{natbib}  % DO NOT CHANGE THIS
\usepackage{caption}  % DO NOT CHANGE THIS
\frenchspacing % DO NOT CHANGE THIS
\setlength{\pdfpagewidth}{8.5in} % DO NOT CHANGE THIS
\setlength{\pdfpageheight}{11in} % DO NOT CHANGE THIS
%
% Keep the \pdfinfo as shown here. There's no need
% for you to add the /Title and /Author tags.
\pdfinfo{
/TemplateVersion (2026.1)
}
```

Preparing Your Paper

After the preamble above, you should prepare your paper as follows:

```
\begin{document}
\maketitle
\begin{abstract}
%...
\end{abstract}
```

If you want to add links to the paper's code, dataset(s), and extended version or similar this is the place to add them, within a *links* environment:

```
\begin{links}
  \link{Code}{https://aaai.org/example/guidelines}
  \link{Datasets}{https://aaai.org/example/datasets}
  \link{Extended version}{https://aaai.org/example}
\end{links}
```

Make sure that you do not de-anonymize yourself with these links.
You should then continue with the body of your paper. Your paper must conclude with the references, which should be inserted as follows:

```
% References and End of Paper
% These lines must be placed at the end of your paper
\bibliography{Bibliography-File}
\end{document}
```

```
\begin{document}\\
\maketitle\\
...\\
\bibliography{Bibliography-File}\\
\end{document}\\
```

Commands and Packages That May Not Be Used

There are a number of packages, commands, scripts, and macros that are incompatable with aaai2026.sty. The common ones are listed in tables 4 and 5. Generally, if a command, package, script, or macro alters floats, margins, fonts, sizing, linespacing, or the presentation of the references and citations, it is unacceptable. Note that negative vskip and vspace may not be used except in certain rare occurances, and may never be used around tables, figures, captions, sections, subsections, subsubsections, or references.

Page Breaks

For your final camera ready copy, you must not use any page break commands. References must flow directly after the text without breaks. Note that some conferences require references to be on a separate page during the review process. AAAI Press, however, does not require this condition for the final paper.

Paper Size, Margins, and Column Width

Papers must be formatted to print in two-column format on 8.5 x 11 inch US letter-sized paper. The margins must be exactly as follows:

- Top margin: 1.25 inches (first page), .75 inches (others)
- Left margin: .75 inches
- Right margin: .75 inches
- Bottom margin: 1.25 inches

The default paper size in most installations of LaTeX is A4. However, because we require that your electronic paper be formatted in US letter size, the preamble we have provided includes commands that alter the default to US letter size. Please note that using any other package to alter page size (such as, but not limited to the Geometry package) will result in your final paper being returned to you for correction.

Column Width and Margins. To ensure maximum readability, your paper must include two columns. Each column should be 3.3 inches wide (slightly more than 3.25 inches), with a .375 inch (.952 cm) gutter of white space between the two columns. The aaai2026.sty file will automatically create these columns for you.

\abovecaption	\abovedisplay	\addevensidemargin	\addsidemargin
\addtolength	\baselinestretch	\belowcaption	\belowdisplay
\break	\clearpage	\clip	\columnsep
\float	\input	\input	\linespread
\newpage	\pagebreak	\renewcommand	\setlength
\text height	\tiny	\top margin	\trim
\vskip{-	\vspace{-		

Table 4: Commands that must not be used

authblk	babel	cjk	dvips
epsf	epsfig	euler	float
fullpage	geometry	graphics	hyperref
layout	linespread	lmodern	maltepaper
navigator	pdfcomment	pgfplots	psfig
pstricks	t1enc	titlesec	tocbind
ulem			

Table 5: LaTeX style packages that must not be used.

Overlength Papers

If your paper is too long and you resort to formatting tricks to make it fit, it is quite likely that it will be returned to you. The best way to retain readability if the paper is overlength is to cut text, figures, or tables. There are a few acceptable ways to reduce paper size that don't affect readability. First, turn on \frenchspacing, which will reduce the space after periods. Next, move all your figures and tables to the top of the page. Consider removing less important portions of a figure. If you use \centering instead of \begin{center} in your figure environment, you can also buy some space. For mathematical environments, you may reduce fontsize **but not below 6.5 point**.

Commands that alter page layout are forbidden. These include \columnsep, \float, \topmargin, \topskip, \textheight, \textwidth, \oddsidemargin, and \evensizemargin (this list is not exhaustive). If you alter page layout, you will be required to pay the page fee. Other commands that are questionable and may cause your paper to be rejected include \parindent, and \parskip. Commands that alter the space between sections are forbidden. The title sec package is not allowed. Regardless of the above, if your paper is obviously "squeezed" it is not going to to be accepted. Options for reducing the length of a paper include reducing the size of your graphics, cutting text, or paying the extra page charge (if it is offered).

Type Font and Size

Your paper must be formatted in Times Roman or Nimbus. We will not accept papers formatted using Computer Modern or Palatino or some other font as the text or heading typeface. Sans serif, when used, should be Courier. Use Symbol or Lucida or Computer Modern for *mathematics only.*

Do not use type 3 fonts for any portion of your paper, including graphics. Type 3 bitmapped fonts are designed for fixed resolution printers. Most print at 300 dpi even if the printer resolution is 1200 dpi or higher. They also often cause high resolution imagesetter devices to crash. Consequently, AAAI will not accept electronic files containing obsolete type 3 fonts. Files containing those fonts (even in graphics) will be rejected. (Authors using blackboard symbols must avoid packages that use type 3 fonts.)

Fortunately, there are effective workarounds that will prevent your file from embedding type 3 bitmapped fonts. The easiest workaround is to use the required times, helvet, and courier packages with LaTeX2e. (Note that papers formatted in this way will still use Computer Modern for the mathematics. To make the math look good, you'll either have to use Symbol or Lucida, or you will need to install type 1 Computer Modern fonts — for more on these fonts, see the section "Obtaining Type 1 Computer Modern.")

If you are unsure if your paper contains type 3 fonts, view the PDF in Acrobat Reader. The Properties/Fonts window will display the font name, font type, and encoding properties of all the fonts in the document. If you are unsure if your graphics contain type 3 fonts (and they are PostScript or encapsulated PostScript documents), create PDF versions of them, and consult the properties window in Acrobat Reader.

The default size for your type must be ten-point with twelve-point leading (line spacing). Start all pages (except the first) directly under the top margin. (See the next section for instructions on formatting the title page.) Indent ten points when beginning a new paragraph, unless the paragraph begins directly below a heading or subheading.

Obtaining Type 1 Computer Modern for LaTeX. If you use Computer Modern for the mathematics in your paper (you cannot use it for the text) you may need to download type 1 Computer fonts. They are available without charge from the American Mathematical Society: http://www.ams.org/tex/type1-fonts.html.

Nonroman Fonts. If your paper includes symbols in other languages (such as, but not limited to, Arabic, Chinese, Hebrew, Japanese, Thai, Russian and other Cyrillic languages), you must restrict their use to bit-mapped figures.

Title and Authors

Your title must appear centered over both text columns in sixteen-point bold type (twenty-four point leading). The title must be written in Title Case according to the Chicago Manual of Style rules. The rules are a bit involved, but in general verbs (including short verbs like be, is, using, and go), nouns, adverbs, adjectives, and pronouns should be capitalized, (including both words in hyphenated terms), while articles, conjunctions, and prepositions are lower case unless they directly follow a colon or long dash. You can use the online tool https://titlecaseconverter.com/ to double-check the proper capitalization (select the ”Chicago” style and mark the ”Show explanations” checkbox).

Author’s names should appear below the title of the paper, centered in twelve-point type (with fifteen point leading), along with affiliation(s) and complete address(es) (including electronic mail address if available) in nine-point roman type (the twelve point leading). You should begin the two-column format when you come to the abstract.

Formatting Author Information. Author information has to be set according to the following specification depending if you have one or more than one affiliation. You may not use a table nor may you employ the \authorblk.sty package. For one or several authors from the same institution, please separate them with commas and write all affiliation directly below (one affiliation per line) using the macros \author and \affiliations:

```
\author{
    Author 1, ..., Author n\\
}
\affiliations {
    Address line\\
    ... \\
    Address line\\
}
```

For authors from different institutions, use \textsuperscript {\rm x } to match authors and affiliations. Notice that there should not be any spaces between the author name (or comma following it) and the superscript.

```
\author{
    AuthorOne\equalcontrib\textsuperscript{\rm 1,\rm 2},
    AuthorTwo\equalcontrib\textsuperscript{\rm 2},\\
    AuthorThree\textsuperscript{\rm 3},\\
    AuthorFour\textsuperscript{\rm 4},
    AuthorFive \textsuperscript{\rm 5}}
}
\affiliations {
    \textsuperscript{\rm 1}AffiliationOne,\\
    \textsuperscript{\rm 2}AffiliationTwo,\\
    \textsuperscript{\rm 3}AffiliationThree,\\
    \textsuperscript{\rm 4}AffiliationFour,\\
    \textsuperscript{\rm 5}AffiliationFive\\
    \{email, email\}@affiliation.com,
    email@affiliation.com,
    email@affiliation.com,
    email@affiliation.com
}
```

You can indicate that some authors contributed equally using the \equalcontrib command. This will add a marker after the author names and a footnote on the first page.

Note that you may want to break the author list for better visualization. You can achieve this using a simple line break (\\).

LaTeX Copyright Notice

The copyright notice automatically appears if you use aaai2026.sty. It has been hardcoded and may not be disabled.

Credits

Any credits to a sponsoring agency should appear in the acknowledgments section, unless the agency requires different placement. If it is necessary to include this information on the front page, use \thanks in either the \author or \title commands. For example:

```
\title{Very Important Results in AI\thanks{This work is
supported by everybody.}}
```

Multiple \thanks commands can be given. Each will result in a separate footnote indication in the author or title with the corresponding text at the botton of the first column of the document. Note that the \thanks command is fragile. You will need to use \protect.

Please do not include \pubnote commands in your document.

Abstract

Follow the example commands in this document for creation of your abstract. The command \begin{abstract} will automatically indent the text block. Please do not indent it further. Do not include references in your abstract!

Page Numbers

Do not print any page numbers on your paper. The use of \pagestyle is forbidden.

Text

The main body of the paper must be formatted in black, ten-point Times Roman with twelve-point leading (line spacing). You may not reduce font size or the linespacing. Commands that alter font size or line spacing (including, but not limited to baselinestretch, baselineshift, linespread, and others) are expressly forbidden. In addition, you may not use color in the text.

Citations

Citations within the text should include the author’s last name and year, for example (Newell 1980). Append lowercase letters to the year in cases of ambiguity. Multiple authors should be treated as follows: (Feigenbaum and Engelmore 1988) or (Ford, Hayes, and Glymour 1992). In the case of four or more authors, list only the first author, followed by et al. (Ford et al. 1997).

Extracts

Long quotations and extracts should be indented ten points from the left and right margins.

> This is an example of an extract or quotation. Note the indent on both sides. Quotation marks are not necessary if you offset the text in a block like this, and properly identify and cite the quotation in the text.

Footnotes

Use footnotes judiciously, taking into account that they interrupt the reading of the text. When required, they should be consecutively numbered throughout with superscript Arabic numbers. Footnotes should appear at the bottom of the page, separated from the text by a blank line space and a thin, half-point rule.

Headings and Sections

When necessary, headings should be used to separate major sections of your paper. Remember, you are writing a short paper, not a lengthy book! An overabundance of headings will tend to make your paper look more like an outline than a paper. The aaai2026.sty package will create headings for you. Do not alter their size nor their spacing above or below.

Section Numbers. The use of section numbers in AAAI Press papers is optional. To use section numbers in LaTeX, uncomment the setcounter line in your document preamble and change the 0 to a 1. Section numbers should not be used in short poster papers and/or extended abstracts.

Section Headings. Sections should be arranged and headed as follows:

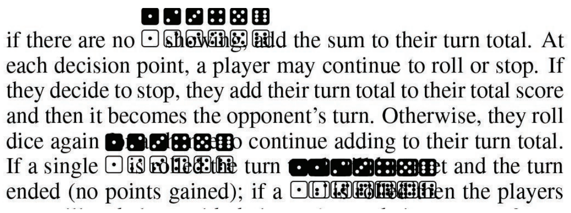
1. Main content sections
2. Appendices (optional)
3. Ethical Statement (optional, unnumbered)
4. Acknowledgements (optional, unnumbered)
5. References (unnumbered)

Appendices. Any appendices must appear after the main content. If your main sections are numbered, appendix sections must use letters instead of arabic numerals. In LaTeX you can use the `\appendix` command to achieve this effect and then use `\section{Heading}` normally for your appendix sections.

Ethical Statement. You can write a statement about the potential ethical impact of your work, including its broad societal implications, both positive and negative. If included, such statement must be written in an unnumbered section titled *Ethical Statement*.

Acknowledgments. The acknowledgments section, if included, appears right before the references and is headed "Acknowledgments". It must not be numbered even if other sections are (use `\section*{Acknowledgements}` in LaTeX). This section includes acknowledgments of help from associates and colleagues, credits to sponsoring agencies, financial support, and permission to publish. Please acknowledge other contributors, grant support, and so forth, in this section. Do not put acknowledgments in a footnote on the first page. If your grant agency requires acknowledgment of the grant on page 1, limit the footnote to the required statement, and put the remaining acknowledgments at the back. Please try to limit acknowledgments to no more than three sentences.

References. The references section should be labeled "References" and must appear at the very end of the paper (don't end the paper with references, and then put a figure by itself on the last page). A sample list of references is given later on in these instructions. Please use a consistent format for references. Poorly prepared or sloppy references reflect badly on the quality of your paper and your research. Please prepare complete and accurate citations.



if there are no ⚀ showing, add the sum to their turn total. At each decision point, a player may continue to roll or stop. If they decide to stop, they add their turn total to their total score and then it becomes the opponent's turn. Otherwise, they roll dice again [illegible] continue adding to their turn total. If a single ⚀ is rolled the turn [illegible]et and the turn ended (no points gained); if a ⚀⚁ is rolled then the players

Figure 2: Using the trim and clip commands produces fragile layers that can result in disasters (like this one from an actual paper) when the color space is corrected or the PDF combined with others for the final proceedings. Crop your figures properly in a graphics program – not in LaTeX.

Illustrations and Figures

Your paper must compile in PDFLaTeX. Consequently, all your figures must be .jpg, .png, or .pdf. You may not use the .gif (the resolution is too low), .ps, or .eps file format for your figures.

Figures, drawings, tables, and photographs should be placed throughout the paper on the page (or the subsequent page) where they are first discussed. Do not group them together at the end of the paper. If placed at the top of the paper, illustrations may run across both columns. Figures must not invade the top, bottom, or side margin areas. Figures must be inserted using the \usepackage{graphicx}. Number figures sequentially, for example, figure 1, and so on. Do not use minipage to group figures.

If you normally create your figures using pgfplots, please create the figures first, and then import them as pdfs with proper bounding boxes, as the bounding and trim boxes created by pfgplots are fragile and not valid.

When you include your figures, you must crop them **outside** of LaTeX. The command \includegraphics*[clip=true, viewport 0 0 10 10]... might result in a PDF that looks great, but the image is **not really cropped.** The full image can reappear (and obscure whatever it is overlapping) when page numbers are applied or color space is standardized. Figures 2, and 3 display some unwanted results that often occur.

If your paper includes illustrations that are not compatible with PDFTeX (such as .eps or .ps documents), you will need to convert them. The epstopdf package will usually work for eps files. You will need to convert your ps files to PDF in either case.

Figure Captions. The illustration number and caption must appear *under* the illustration. Labels and other text with the actual illustration must be at least nine-point type. However, the font and size of figure captions must be 10 point roman. Do not make them smaller, bold, or italic. (Individual words may be italicized if the context requires differentiation.)

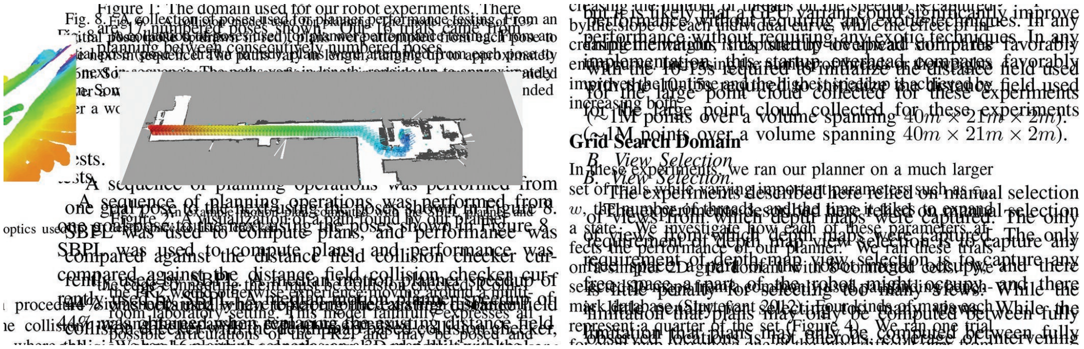


Figure 3: Adjusting the bounding box instead of actually removing the unwanted data resulted multiple layers in this paper. It also needlessly increased the PDF size. In this case, the size of the unwanted layer doubled the paper's size, and produced the following surprising results in final production. Crop your figures properly in a graphics program. Don't just alter the bounding box.

Tables

Tables

Tables should be presented in 10 point roman type. If necessary, they may be altered to 9 point type. You must not use `\resizebox` or other commands that resize the entire table to make it smaller, because you can't control the final font size this way. If your table is too large you can use `\setlength{\tabcolsep}{1mm}` to compress the columns a bit or you can adapt the content (e.g.: reduce the decimal precision when presenting numbers, use shortened column titles, make some column duble-line to get it narrower).

Tables that do not fit in a single column must be placed across double columns. If your table won't fit within the margins even when spanning both columns and using the above techniques, you must split it in two separate tables.

Table Captions. The number and caption for your table must appear *under* (not above) the table. Additionally, the font and size of table captions must be 10 point roman and must be placed beneath the figure. Do not make them smaller, bold, or italic. (Individual words may be italicized if the context requires differentiation.)

Low-Resolution Bitmaps. You may not use low-resolution (such as 72 dpi) screen-dumps and GIF files—these files contain so few pixels that they are always blurry, and illegible when printed. If they are color, they will become an indecipherable mess when converted to black and white. This is always the case with gif files, which should never be used. The resolution of screen dumps can be increased by reducing the print size of the original file while retaining the same number of pixels. You can also enlarge files by manipulating them in software such as PhotoShop. Your figures should be 300 dpi when incorporated into your document.

LaTeX Overflow. LaTeX users please beware: LaTeX will sometimes put portions of the figure or table or an equation in the margin. If this happens, you need to make the figure or table span both columns. If absolutely necessary, you may reduce the figure, or reformat the equation, or reconfigure the table. **Check your log file!** You must fix any overflow into the margin (that means no overfull boxes in LaTeX). **Nothing is permitted to intrude into the margin or gutter.**

Using Color. Use of color is restricted to figures only. It must be WACG 2.0 compliant. (That is, the contrast ratio must be greater than 4.5:1 no matter the font size.) It must be CMYK, NOT RGB. It may never be used for any portion of the text of your paper. The archival version of your paper will be printed in black and white and grayscale. The web version must be readable by persons with disabilities. Consequently, because conversion to grayscale can cause undesirable effects (red changes to black, yellow can disappear, and so forth), we strongly suggest you avoid placing color figures in your document. If you do include color figures, you must (1) use the CMYK (not RGB) colorspace and (2) be mindful of readers who may happen to have trouble distinguishing colors. Your paper must be decipherable without using color for distinction.

Drawings. We suggest you use computer drawing software (such as Adobe Illustrator or, (if unavoidable), the drawing tools in Microsoft Word) to create your illustrations. Do not use Microsoft Publisher. These illustrations will look best if all line widths are uniform (half- to two-point in size), and you do not create labels over shaded areas. Shading should be 133 lines per inch if possible. Use Times Roman or Helvetica for all figure call-outs. **Do not use hairline width lines** — be sure that the stroke width of all lines is at least .5 pt. Zero point lines will print on a laser printer, but will completely disappear on the high-resolution devices used by our printers.

Photographs and Images. Photographs and other images should be in grayscale (color photographs will not reproduce well; for example, red tones will reproduce as black, yellow may turn to white, and so forth) and set to a minimum of 300 dpi. Do not prescreen images.

Resizing Graphics. Resize your graphics **before** you include them with LaTeX. You may **not** use trim or clip op-

Algorithm 1: Example algorithm

```
Input: Your algorithm's input
Parameter: Optional list of parameters
Output: Your algorithm's output
Let $t = 0$.
while condition do
    Do some action.
    if conditional then
        Perform task A.
    else
        Perform task B.
    end if
end while
return solution
```

Listing 1: Example listing `quicksort.hs`

```haskell
quicksort :: Ord a => [a] -> [a]
quicksort []     = []
quicksort (p:xs) = (quicksort lesser) ++
     [p] ++ (quicksort greater)
  where
    lesser  = filter (< p) xs
    greater = filter (>= p) xs
```

tions as part of your \includegraphics command. Resize the media box of your PDF using a graphics program instead.

Fonts in Your Illustrations. You must embed all fonts in your graphics before including them in your LaTeX document.

Algorithms. Algorithms and/or programs are a special kind of figures. Like all illustrations, they should appear floated to the top (preferably) or bottom of the page. However, their caption should appear in the header, left-justified and enclosed between horizontal lines, as shown in Algorithm 1. The algorithm body should be terminated with another horizontal line. It is up to the authors to decide whether to show line numbers or not, how to format comments, etc.

In LaTeX algorithms may be typeset using the `algorithm` and `algorithmic` packages, but you can also use one of the many other packages for the task.

Listings. Listings are much like algorithms and programs. They should also appear floated to the top (preferably) or bottom of the page. Listing captions should appear in the header, left-justified and enclosed between horizontal lines as shown in Listing 1. Terminate the body with another horizontal line and avoid any background color. Line numbers, if included, must appear within the text column.

References

The AAAI style includes a set of definitions for use in formatting references with BibTeX. These definitions make the bibliography style fairly close to the ones specified in the Reference Examples appendix below. To use these definitions, you also need the BibTeX style file "aaai2026.bst," available in the AAAI Author Kit on the AAAI web site. Then, at the end of your paper but before \enddocument, you need to put the following lines:

\bibliography{bibfile1,bibfile2,...}

Please note that the aaai2026.sty class already sets the bibliographystyle for you, so you do not have to place any \bibliographystyle command in the document yourselves. The aaai2026.sty file is incompatible with the hyperref and navigator packages. If you use either, your references will be garbled and your paper will be returned to you.

References may be the same size as surrounding text. However, in this section (only), you may reduce the size to *small* (9pt) if your paper exceeds the allowable number of pages. Making it any smaller than 9 point with 10 point linespacing, however, is not allowed.

The list of files in the \bibliography command should be the names of your BibTeX source files (that is, the .bib files referenced in your paper).

The following commands are available for your use in citing references:

cite: Cites the given reference(s) with a full citation. This appears as "(Author Year)" for one reference, or "(Author Year; Author Year)" for multiple references.

shortcite: Cites the given reference(s) with just the year. This appears as "(Year)" for one reference, or "(Year; Year)" for multiple references.

citeauthor: Cites the given reference(s) with just the author name(s) and no parentheses.

citeyear: Cites the given reference(s) with just the date(s) and no parentheses.

You may also use any of the *natbib* citation commands.

Proofreading Your PDF

Please check all the pages of your PDF file. The most commonly forgotten element is the acknowledgements — especially the correct grant number. Authors also commonly forget to add the metadata to the source, use the wrong reference style file, or don't follow the capitalization rules or comma placement for their author-title information properly. A final common problem is text (expecially equations) that runs into the margin. You will need to fix these common errors before submitting your file.

Improperly Formatted Files

In the past, AAAI has corrected improperly formatted files submitted by the authors. Unfortunately, this has become an increasingly burdensome expense that we can no longer absorb). Consequently, if your file is improperly formatted, it will be returned to you for correction.

Naming Your Electronic File

We require that you name your LaTeX source file with the last name (family name) of the first author so that it can easily be differentiated from other submissions. Complete file-naming instructions will be provided to you in the submission instructions.

Submitting Your Electronic Files to AAAI

Instructions on paper submittal will be provided to you in your acceptance letter.

Inquiries

If you have any questions about the preparation or submission of your paper as instructed in this document, please contact AAAI Press at the address given below. If you have technical questions about implementation of the aaai style file, please contact an expert at your site. We do not provide technical support for LaTeX or any other software package. To avoid problems, please keep your paper simple, and do not incorporate complicated macros and style files.

AAAI Press
1101 Pennsylvania Ave, NW Suite 300
Washington, DC 20004 USA
Telephone: 1-202-360-4062
E-mail: See the submission instructions for your particular conference or event.

Additional Resources

LaTeX is a difficult program to master. If you've used that software, and this document didn't help or some items were not explained clearly, we recommend you read Michael Shell's excellent document (testflow doc.txt V1.0a 2002/08/13) about obtaining correct PS/PDF output on LaTeX systems. (It was written for another purpose, but it has general application as well). It is available at www.ctan.org in the tex-archive.

Reference Examples

Formatted bibliographies should look like the following examples. You should use BibTeX to generate the references. Missing fields are unacceptable when compiling references, and usually indicate that you are using the wrong type of entry (BibTeX class).

Book with multiple authors Use the `@book` class.

.

Journal and magazine articles Use the `@article` class.

.

.

Proceedings paper published by a society, press or publisher Use the `@inproceedings` class. You may abbreviate the *booktitle* field, but make sure that the conference edition is clear.

.

.

University technical report Use the `@techreport` class.

.

Dissertation or thesis Use the `@phdthesis` class.

.

Forthcoming publication Use the `@misc` class with a `note="Forthcoming"` annotation.

```
@misc(key,
  [...]
  note="Forthcoming",
)
```

.

ArXiv paper Fetch the BibTeX entry from the "Export Bibtex Citation" link in the arXiv website. Notice it uses the `@misc` class instead of the `@article` one, and that it includes the `eprint` and `archivePrefix` keys.

```
@misc(key,
  [...]
  eprint="xxxx.yyyy",
  archivePrefix="arXiv",
)
```

.

Website or online resource Use the `@misc` class. Add the url in the `howpublished` field and the date of access in the `note` field:

```
@misc(key,
  [...]
  howpublished="\url{http://...}",
  note="Accessed: YYYY-mm-dd",
)
```

.

For the most up to date version of the AAAI reference style, please consult the *AI Magazine* Author Guidelines at https://aaai.org/ojs/index.php/aimagazine/about/submissions#authorGuidelines

Acknowledgments

AAAI is especially grateful to Peter Patel Schneider for his work in implementing the original aaai.sty file, liberally using the ideas of other style hackers, including Barbara Beeton. We also acknowledge with thanks the work of George Ferguson for his guide to using the style and BibTeX files — which has been incorporated into this document — and Hans Guesgen, who provided several timely modifications, as well as the many others who have, from time to time, sent in suggestions on improvements to the AAAI style. We are especially grateful to Francisco Cruz, Marc Pujol-Gonzalez, and Mico Loretan for the improvements to the BibTeX and LaTeX files made in 2020.

The preparation of the LaTeX and BibTeX files that implement these instructions was supported by Schlumberger Palo Alto Research, AT&T Bell Laboratories, Morgan Kaufmann Publishers, The Live Oak Press, LLC, and AAAI Press. Bibliography style changes were added by Sunil Issar. \pubnote was added by J. Scott Penberthy. George Ferguson added support for printing the AAAI copyright slug. Additional changes to aaai2026.sty and aaai2026.bst have been made by Francisco Cruz and Marc Pujol-Gonzalez.

Thank you for reading these instructions carefully. We look forward to receiving your electronic files!

References

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. SPICE: Semantic Propositional Image Caption Evaluation. In *Proceedings of the 14th European Conference on Computer Vision (ECCV 2016)*, 382–398. Cham, Switzerland: Springer.

Ataallah, K.; Gou, C.; Abdelrahman, E.; Pahwa, K.; Ding, J.; and Elhoseiny, M. 2025. InfiniBench: A Comprehensive Benchmark for Large Multimodal Models in Very Long Video Understanding. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR-2025)*. Singapore.

Ataallah, K.; Shen, X.; Abdelrahman, E.; Sleiman, E.; Zhuge, M.; Ding, J.; Zhu, D.; Schmidhuber, J.; and Elhoseiny, M. 2024. Goldfish: Vision-Language Understanding of Arbitrarily Long Videos. In *Proceedings of the European Conference on Computer Vision (ECCV 2024)*, 251–267. Cham, Switzerland: Springer.

Banerjee, S.; and Lavie, A. 2005. 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*, 65–72. Ann Arbor, Mich.: Association for Computational Linguistics.

Chan, D. M.; Petryk, S.; Gonzalez, J. E.; Darrell, T.; and Canny, J. 2023. CLAIR: Evaluating Image Captions with Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP-2023)*, 13638–13646. Singapore: Association for Computational Linguistics.

Chen, D. L.; and Dolan, W. B. 2011. Collecting Highly Parallel Data for Paraphrase Evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*. Portland, OR.

Chen, Y.; Xue, F.; Li, D.; Hu, Q.; Zhu, L.; Li, X.; Fang, Y.; Tang, H.; Yang, S.; Liu, Z.; He, Y.; Yin, H.; Molchanov, P.; Kautz, J.; Fan, L.; Zhu, Y.; Lu, Y.; and Han, S. 2025. LongVILA: Scaling Long-Context Visual Language Models for Long Videos. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR-2025)*. Singapore: ICLR.

Cheng, K.; Song, W.; Fan, J.; Ma, Z.; Sun, Q.; Xu, F.; Yan, C.; Chen, N.; Zhang, J.; and Chen, J. 2025. CapArena: Benchmarking and Analyzing Detailed Image Captioning in the LLM Era. ArXiv preprint, arXiv:2503.12329.

Choi, C.; Kim, J.; Lee, S.; Kwon, J.; Gu, S.; Kim, Y.; Cho, M.; and Sohn, J.-y. 2024. Linq-Embed-Mistral: Technical Report. arXiv:2412.03223.

Dubey, H.; and Pack, C. 2025. Leveraging Textual Memory and Key Frame Reasoning for Full Video Understanding Using Off-the-Shelf LLMs and VLMs (Student Abstract). In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25)*, 12445–12446. Menlo Park, Calif.: AAAI Press.

Frohmann, M.; Sterner, I.; Vulić, I.; Minixhofer, B.; and Schedl, M. 2024. Segment Any Text: A Universal Approach for Robust, Efficient and Adaptable Sentence Segmentation. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 11908–11941. Miami, Florida, USA: Association for Computational Linguistics.

Lee, C.; Roy, R.; Xu, M.; Raiman, J.; Shoeybi, M.; Catanzaro, B.; and Ping, W. 2024. NV-Embed: Improved Techniques for Training LLMs as Generalist Embedding Models. arXiv:2405.17428.

Li, B.; Zhu, L.; Tian, R.; Tan, S.; Chen, Y.; Lu, Y.; Cui, Y.; Veer, S.; Ehrlich, M.; Philion, J.; Weng, X.; Xue, F.; Fan, J.; Zhu, Y.; Kautz, J.; Tao, A.; Liu, M.-Y.; Fidler, S.; Ivanovic, B.; Darrell, T.; Malik, J.; Han, S.; and Pavone, M. 2024. Wolf: Dense Video Captioning with a World Summarization Framework. ArXiv preprint, arXiv:2407.18908.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Meng, R.; Liu, Y.; Joty, S. R.; Xiong, C.; Zhou, Y.; and Yavuz, S. 2024. SFR-Embedding-2: Advanced Text Embedding with Multi-stage Training. https://huggingface.co/Salesforce/SFR-Embedding-2_R.

Nagrani, A.; Zhang, M.; Mehran, R.; Hornung, R.; Gundavarapu, N. B.; Jha, N.; Myers, A.; Zhou, X.; Gong, B.; Schmid, C.; Sirotenko, M.; Zhu, Y.; and Weyand, T. 2025. Neptune: The Long Orbit to Benchmarking Long Video Understanding. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR-2025)*. Singapore.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pa.: Association for Computational Linguistics.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. Hong Kong, China: Association for Computational Linguistics.

Rohrbach, A.; Rohrbach, M.; Tandon, N.; and Schiele, B. 2015. A Dataset for Movie Description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society.

Sarto, S.; Barraco, M.; Cornia, M.; Baraldi, L.; and Cucchiara, R. 2023. Positive-Augmented Contrastive Learning for Image and Video Captioning Evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pat-*

tern Recognition (CVPR 2023), 6914–6924. Los Alamitos, Calif.: IEEE Computer Society.

Shen, X.; Xiong, Y.; Zhao, C.; Wu, L.; Chen, J.; Zhu, C.; Liu, Z.; Xiao, F.; Varadarajan, B.; Bordes, F.; Liu, Z.; Xu, H.; Kim, H. J.; Soran, B.; Krishnamoorthi, R.; Elhoseiny, M.; and Chandra, V. 2025. LongVU: Spatiotemporal Adaptive Compression for Long Video-Language Understanding. In *Proceedings of the 42nd International Conference on Machine Learning (ICML-25)*. Vancouver, Canada: PMLR.

Shi, Y.; Yang, X.; Xu, H.; Yuan, C.; Li, B.; Hu, W.; and Zha, Z.-J. 2022. 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 (CVPR 2022)*, 17929–17938. Los Alamitos, Calif.: IEEE Computer Society.

Vedantam, R.; Zitnick, C. L.; and Parikh, D. 2015. CIDEr: Consensus-Based Image Description Evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, 4566–4575. Washington, D.C.: IEEE Computer Society.

Wang, J.; Yuan, L.; Zhang, Y.; and Sun, H. 2024. Tarsier: Recipes for Training and Evaluating Large Video Description Models. arXiv:2407.00634.

Wu, H.; Li, D.; Chen, B.; and Li, J. 2024. LongVideoBench: A Benchmark for Long-context Interleaved Video-Language Understanding. In *Advances in Neural Information Processing Systems 37 (NeurIPS-2024), Datasets and Benchmarks Track*. San Diego, CA: Curran Associates.

Yuan, L.; Wang, J.; Sun, H.; Zhang, Y.; and Lin, Y. 2025. Tarsier2: Advancing Large Vision-Language Models from Detailed Video Description to Comprehensive Video Understanding. arXiv:2501.07888.

Zhang, D.; Li, J.; Zeng, Z.; and Wang, F. 2024. Jasper and Stella: distillation of SOTA embedding models. arXiv:2412.19048.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020)*. Addis Ababa, Ethiopia: OpenReview.net.

Zhou, L.; Xu, C.; and Corso, J. J. 2018. Towards Automatic Learning of Procedures From Web Instructional Videos. In *AAAI Conference on Artificial Intelligence*, 7590–7598.