rigorous.

AI Review Report

For the manuscript

Iterative Proposal Refinement for Weakly-Supervised Video Grounding

Publication Outlets: CVPR
Review Focus:

Thank you for using the Rigorous AI Reviewer!

We're dedicated to providing actionable, high-quality feedback that accelerates your revision process and boosts your chances of publication. To help us improve the system, please consider completing our short feedback survey. Your input directly contributes to making this tool more useful, accurate, and impactful for the research community. All responses are confidential and sincerely appreciated.

Feedback Link: Feedback Form

Important Note: Like real human reviews, this AI-generated feedback may occasionally include hallucinations, overconfident statements, vague suggestions, or simply a false statement. Still, we hope you find it insightful and helpful in improving your manuscript for publication. This is very much an MVP (Minimum Viable Product) and is far from being the best version it can be. We are committed to making it truly excellent, and your feedback is essential to help us get there.

Want to submit more manuscripts for review?

https://www.rigorous.company
...We process new submissions for free upon receiving your feedback

Want to help improve the AI Reviewer?

https://github.com/robertjakob/rigorous
...Give us a Star to stay up to date on future improvements and new features of the AI Reviewer!

Executive Summary

This manuscript introduces IRON, an Iterative Proposal Refinement network designed to advance weakly-supervised video grounding (WSVG). The proposed framework addresses the challenge of localizing events in untrimmed videos using only video-level annotations, a task complicated by the absence of explicit temporal supervision and the complexity of multi-event queries. IRON distinguishes itself through dual lightweight distillation branches that model cross-modal correspondence at both semantic and conceptual levels, coupled with an iterative label propagation strategy to enhance event coverage during proposal refinement. Extensive experiments and ablation studies on benchmark datasets such as Charades-STA and ActivityNet Captions demonstrate the method's effectiveness, with IRON achieving state-of-the-art performance and offering promising directions for practical applications in video retrieval and content understanding. The manuscript's strengths lie in its clear motivation, methodological innovation, and comprehensive experimental validation. The integration of cross-modal distillation and iterative refinement is conceptually sound and empirically effective, and the paper is generally well-structured and accessible to the CVPR audience. However, several weaknesses limit its impact: the technical details of key modules, particularly the distillation branches and label propagation mechanism, are insufficiently described, hindering reproducibility and assessment of novelty. The abstract and introduction lack explicit quantitative results and clear research questions, while the discussion and conclusion do not adequately address limitations, scalability, or broader societal implications. Additionally, the manuscript would benefit from improved clarity, more consistent terminology, and enhanced statistical rigor in presenting results. To strengthen the manuscript, the authors should provide more detailed descriptions of the proposed modules and clarify the rationale behind key design choices. Explicitly reporting quantitative improvements over baselines, justifying hyperparameter selections, and incorporating statistical significance testing will enhance scientific rigor. The paper should also include a dedicated keywords section, improve narrative flow and figure clarity, and discuss ethical considerations, dataset diversity, and potential limitations or failure cases. Addressing these points will improve the work's clarity, reproducibility, and impact, making it more compelling for both expert and broader CVPR audiences.

Category	Sub-Category
Section Assessment	S1 - Title and Keywords
	S2 - Abstract
	S3 - Introduction
	S4 - Literature Review
	S5 - Methodology
	S6 - Results
	S7 - Discussion
	S8 - Conclusion
	S9 - References
	S10 - Supplementary Materials
Rigor Assessment	R1 - Originality and Contribution
	R2 - Impact and Significance
	R3 - Ethics and Compliance
	R4 - Data and Code Availability
	R5 - Statistical Rigor
	R6 - Technical Accuracy
	R7 - Consistency
Writing Assessment	W1 - Language and Style
	W2 - Narrative and Structure
	W3 - Clarity and Conciseness
	W4 - Terminology Consistency
	W5 - Inclusive Language
	W6 - Citation Formatting
	W7 - Target Audience Alignment

S1 - Title and Keywords

The manuscript's title, 'Iterative Proposal Refinement for Weakly-Supervised Video Grounding,' is clear and descriptive, effectively conveying the research focus and methodology. However, it uses technical jargon that may limit accessibility to a broader audience and could be made more concise for better discoverability. The absence of a dedicated keywords section is a significant oversight, as it reduces the manuscript's visibility in search engines and academic databases. Overall, the title aligns with field standards, but the manuscript would benefit from keyword optimization and the explicit inclusion of a keywords section.

Remarks	Original	Improved	Explanation
The title is clear but could be optimized for conciseness and broader discoverability by reordering and simplifying terms.	Iterative Proposal Refinement for Weakly-Supervised Video Grounding	Efficient Weakly-Supervised Video Grounding via Iterative Proposal Refinement	This revision maintains the core focus while emphasizing efficiency and improving search engine optimization with key terms.
No dedicated keywords section is present, which limits discoverability.	No keywords provided.	Keywords: Weakly-Supervised Learning, Video Grounding, Proposal Refinement, Vision-Language Models, Temporal Localization	Including a keywords section with relevant terms increases the manuscript's visibility in search engines and academic databases.
The title contains technical jargon that may not be immediately accessible to all readers.	Iterative Proposal Refinement for Weakly-Supervised Video Grounding	Iterative Proposal Refinement for Video Event Localization with Weak Supervision	Replacing specialized terms with more general language can broaden the audience and improve initial engagement.

S2 - Abstract

The abstract presents the motivation, methodology, and broad claims of improved performance for the proposed IRON framework. However, it is densely written, lacks clear structure, and omits specific quantitative results or performance metrics, which are essential for demonstrating the significance of the work. Technical terminology is introduced without sufficient explanation, potentially limiting accessibility for non-experts. Minor typographical inconsistencies are present. Overall, the abstract is promising but would benefit from clearer organization, explicit results, and simplified language.

Remarks	Original	Improved	Explanation
The abstract lacks explicit quantitative results, reducing its persuasive power.	Extensive experiments and ablation studies on two challenging WSVG datasets have attested to the effectiveness of our IRON.	Extensive experiments and ablation studies on Charades-STA and ActivityNet Captions datasets demonstrate that IRON achieves state-of-the-art performance, with improvements of up to X% in R1@0.3 and R2@0.5 metrics compared to previous methods.	Including specific datasets and performance metrics substantiates claims and enhances the impact of the abstract.
Technical terms are introduced without explanation, which may confuse readers.	We set up two lightweight distillation branches to uncover the cross-modal correspondence on both the semantic and conceptual levels.	Our method employs two lightweight distillation branches that explicitly model cross-modal correspondence at both semantic and conceptual levels, enhancing proposal quality.	Clarifying the technical approach and its purpose makes the abstract more accessible to a broader audience.

Remarks	Original	Improved	Explanation
The abstract is densely written and lacks clear sectional separation.	Weakly-Supervised Video Grounding (WSVG) aims to localize events of interest in untrimmed videos with only video-level annotations. To date, most of the state-of-the-art WSVG methods follow a two-stage pipeline...	Background: Weakly-Supervised Video Grounding (WSVG) seeks to localize events in untrimmed videos using only video-level annotations. Problem: Existing methods lack explicit cross-modal correspondence modeling and struggle with complex event coverage. Approach: We propose IRON, an iterative proposal refinement network with semantic and conceptual distillation. Results: IRON achieves state-of-the-art performance on Charades-STA and ActivityNet Captions. Code: Available at https://github.com/mengcaopku/IRON.	Structuring the abstract into clear sections improves readability and helps readers quickly grasp the key contributions.

S3 - Introduction

The introduction provides a broad overview of weakly-supervised video grounding and outlines the motivation for the proposed IRON method. However, it lacks a concise synthesis of the field's evolution, a clear and standalone problem statement, and explicit research questions or hypotheses. The significance of the work is implied rather than directly stated, and the structure is somewhat dense, mixing background, problem, and technical details. Improving clarity, focus, and organization would strengthen the introduction's impact.

Remarks	Original	Improved	Explanation
The problem statement is buried within technical details and lacks clarity.	Despite the recent progress, existing proposal generation methods suffer from two drawbacks: 1) lack of explicit correspondence modeling; and 2) partial coverage of complex events.	However, current proposal generation methods face two key limitations: first, they lack explicit modeling of cross-modal correspondences between video segments and language queries; second, they often fail to comprehensively cover complex, multi-event activities within videos.	Explicitly articulating the research gap clarifies the motivation and necessity for the proposed solution.
The introduction lacks explicit research questions or hypotheses.	While the contributions are listed, the specific research questions or hypotheses guiding the study are not explicitly stated.	Add explicit research questions such as: 'Can iterative proposal refinement effectively address coverage and correspondence issues in weakly-supervised video grounding?' or 'Does distilling semantic and conceptual knowledge improve proposal quality?'	Stating research questions clarifies the scientific inquiry and frames the contributions within testable hypotheses.
The structure is dense, mixing background, problem, and technical overview.	The introduction contains a mixture of background, problem statement, and technical overview, leading to a somewhat dense and less cohesive flow.	Reorganize the introduction into clearer subsections: first, provide a concise background; second, explicitly state the core problems; third, outline the research objectives and contributions; finally, discuss the significance.	A clear structure improves readability and logical flow, guiding the reader through the narrative.

S4 - Literature Review

The literature review demonstrates a solid grasp of recent advances in weakly-supervised video grounding and related areas. However, it lacks discussion of foundational methods, provides only superficial analysis of the proposed method's advantages and limitations, and does not sufficiently synthesize or compare existing approaches. Citations are mostly recent but could be more comprehensive and authoritative. Improving historical context, critical analysis, and synthesis would elevate the review's scholarly depth.

Remarks	Original	Improved	Explanation
Foundational or classical methods are underrepresented, limiting historical context.	The review covers many recent methods and datasets but lacks a comprehensive discussion of classical or foundational approaches in weakly-supervised video grounding, especially older or less-cited methods that could provide context for current advances.	Include a section that discusses foundational methods and early approaches in weakly-supervised video grounding, such as initial MIL-based techniques or classical proposal methods, to provide historical context and highlight the progression of the field.	Adding foundational context enriches the review and situates current advances within the field's evolution.
The analysis of IRON's advantages and limitations over prior work is superficial.	The critical analysis of the proposed IRON method's advantages over existing approaches is somewhat superficial, with limited discussion on potential limitations, failure cases, or the reasons behind observed performance gains.	Expand the discussion to critically evaluate the limitations of IRON, such as potential failure modes, computational complexity, or scenarios where it may underperform, and compare these aspects explicitly with prior methods.	A balanced critique enhances scholarly rigor and provides guidance for future research.
The review lists methods and datasets without sufficient synthesis or comparison.	The review tends to list methods and datasets without sufficiently synthesizing how they connect, contrast, or build upon each other, especially in the context of the proposed method.	Incorporate a synthesis paragraph that explicitly compares different approaches, discusses their relative strengths and weaknesses, and positions the IRON method within this landscape, highlighting how it addresses existing gaps.	Synthesis deepens understanding of the field and clarifies the novelty of the current work.

S5 - Methodology

The methodology section introduces an innovative dual-level correspondence modeling approach and iterative proposal refinement, leveraging pre-trained vision-language models. While the technical design is sound and well-motivated, the rationale for specific choices (e.g., semantic/conceptual levels, fixed thresholds) is not fully justified, and adaptive mechanisms are lacking. The section would benefit from more thorough justification of design decisions, adaptive parameter tuning, and explicit discussion of ethical considerations and potential biases.

Remarks	Original	Improved	Explanation
The rationale for selecting semantic and conceptual levels is not fully explained.	We set up two lightweight distillation branches to uncover the cross-modal correspondence on both the semantic and conceptual levels.	Provide a detailed justification for selecting semantic and conceptual levels for correspondence modeling, possibly supported by theoretical or empirical evidence, to clarify why these levels are most effective for weakly-supervised video grounding.	A clear rationale enhances theoretical rigor and helps readers understand the design choices.
The label propagation algorithm uses a fixed IoU threshold, limiting flexibility.	The label propagation algorithm is crafted with a fixed IoU threshold β=0.6.	Introduce an adaptive thresholding mechanism for IoU, possibly based on dataset statistics or learning, to dynamically adjust the threshold for proposal coverage during label propagation.	Adaptive thresholds improve robustness and generalizability across diverse datasets.
The methodology does not discuss ethical considerations or dataset/model biases.	The manuscript lacks explicit discussion of ethical considerations.	Add a dedicated section discussing ethical issues related to dataset collection, potential biases, privacy concerns, and societal implications of deploying weakly-supervised video grounding models.	Addressing ethics promotes responsible research and societal awareness.

S6 - Results

The results section presents comprehensive quantitative comparisons and ablation studies, demonstrating the effectiveness of the proposed IRON framework. However, the presentation of tables and figures lacks consistent formatting, clear labels, and comprehensive captions, which hampers readability. Statistical significance is not explicitly addressed, undermining the rigor of the reported improvements. Enhanced data visualization, explicit statistical analysis, and clearer thematic organization would improve the section's clarity and scientific credibility.

Remarks	Original	Improved	Explanation
Tables lack significance indicators and clear labels.	Table 1. Comparisons (%) with state-of-the-art methods on Charades-STA dataset.	Table 1. Performance comparison of various methods on Charades-STA dataset (in %). Include significance indicators (e.g., * for p<0.05) to clarify statistical relevance.	Significance indicators and clearer labels improve interpretability and emphasize the robustness of improvements.
Statistical tests are not described, reducing scientific rigor.	The comparison results on Charades-STA and ActivityNet Captions datasets are summarized in Table 1 and Table 3, respectively.	The results, summarized in Tables 1 and 3, show that our IRON method consistently outperforms previous approaches across multiple metrics. Statistical significance testing (e.g., t-test) was conducted to confirm these improvements.	Explicit mention of statistical tests and significance levels increases reader confidence and methodological rigor.
Figures are densely packed and lack comprehensive captions.	Figures are densely packed with information, some with unclear labels or small fonts.	Revise figures to include larger fonts, clearer labels, and concise captions highlighting key insights. For example, in Figure 2, explicitly annotate the distribution differences with statistical significance markers.	Improved visualization clarity helps readers quickly grasp the key findings and supports the narrative effectively.

S7 - Discussion

The discussion section highlights the technical contributions and performance improvements of IRON but lacks depth in interpreting the practical and theoretical significance of the results. It does not critically analyze limitations, failure cases, or dataset biases, and the comparison with existing literature is not sufficiently detailed. The narrative is somewhat disjointed, with abrupt transitions between technical and high-level implications. Structuring the discussion, elaborating on limitations, and connecting findings to broader impacts would enhance its clarity and value.

Remarks	Original	Improved	Explanation
The discussion lacks critical analysis of limitations and failure cases.	The limitations are briefly mentioned, but the discussion does not critically analyze failure cases or dataset biases.	A more detailed discussion of potential limitations—such as failure cases in complex scenes, dataset biases, or scalability issues—would provide a balanced perspective and guide future research directions.	Critical reflection on limitations fosters transparency and helps contextualize the results within broader challenges.
The practical and theoretical implications are not sufficiently elaborated.	The paper does not elaborate on how IRON could influence practical applications or advance theoretical understanding.	Elaborating on how IRON's approach could be integrated into real-world systems or how it advances the theoretical framework of weakly-supervised grounding would highlight its broader impact.	Discussing broader implications increases the perceived significance and relevance of the research.
The discussion is disjointed, lacking clear structure and transitions.	The discussion jumps between technical details and high-level implications without clear transitions.	Structuring the discussion with clear subsections—such as 'Results Significance,' 'Methodological Insights,' and 'Broader Implications'—would improve coherence and readability.	A well-structured discussion enhances clarity and helps readers follow the key messages.

S8 - Conclusion

The conclusion summarizes the main contributions and claims effectiveness for the IRON method but lacks specific quantitative evidence and does not explicitly confirm the fulfillment of all research objectives. The implications for future research and practical applications are only briefly mentioned, and the closing statement is generic. Including concrete results, restating objective fulfillment, and providing a strong closing remark would make the conclusion more convincing and impactful.

Remarks	Original	Improved	Explanation
The conclusion lacks specific quantitative evidence to support its claims.	Extensive experiments and ablation studies on two challenging WSVG datasets have attested to the effectiveness of our IRON.	Extensive experiments and ablation studies on Charades-STA and ActivityNet Captions datasets demonstrate that IRON achieves state-of-the-art performance, with improvements of up to X% in R1@0.3 and R2@0.5 metrics compared to previous methods.	Providing concrete datasets and metrics substantiates the claims and enhances credibility.
The conclusion does not explicitly confirm whether all research objectives were achieved.	We make three contributions in this paper: ...	Our three main contributions are: (1) modeling explicit cross-modal correspondence at semantic and conceptual levels to improve proposal quality; (2) developing a label propagation strategy to enhance coverage and reduce bias; (3) achieving new state-of-the-art results on multiple datasets, confirming the effectiveness of our approach.	Restating objectives and linking them to outcomes clarifies the scope of success.
The closing statement is generic and lacks impact.	The extensive experiments attest to the effectiveness of our IRON.	In conclusion, IRON represents a significant step forward in weakly-supervised video grounding, offering both methodological innovations and practical improvements, with promising avenues for future exploration.	A strong closing statement emphasizes the importance and novelty of the work.

S9 - References

The reference list includes many recent and relevant sources but suffers from inconsistent formatting, incomplete citation details, and a lack of logical organization. Some older or less impactful references are included without clear justification, and the absence of a keywords section further limits discoverability. Standardizing citation style, ensuring completeness, and prioritizing recent, high-impact works would enhance the professionalism and scholarly quality of the manuscript.

Remarks	Original	Improved	Explanation
References lack consistent formatting and complete details.	[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision, pages 2425–2433, 2015.	[1] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). VQA: Visual Question Answering. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2425–2433).	Standardizing citation format improves clarity and professionalism, making sources easier to verify.
Older or less impactful references are included without clear justification.	Several references are older (e.g., [10], 1997) or less impactful, with no clear indication of their relevance to recent advances.	Prioritize including recent, high-impact, and highly relevant references, especially recent surveys or seminal works in video-language modeling and weakly-supervised learning, while removing or justifying older or less relevant sources.	Emphasizing recent and relevant works strengthens the manuscript's scholarly foundation.
References are not grouped or organized for easy navigation.	References are ordered numerically but not grouped by theme or publication year, making navigation difficult.	Organize references either alphabetically by author or thematically (e.g., foundational methods, recent advances, datasets). Ensure consistent numbering aligned with in-text citations, and consider grouping related references for clarity.	Logical organization aids reviewers and readers in understanding the scholarly context.

S10 - Supplementary Materials

The supplementary materials provide detailed methodological descriptions, additional experiments, and visualizations that support the main manuscript. However, some sections are overly verbose, figures lack sufficient explanatory labels, and key experimental details are missing or only briefly mentioned. Improving clarity, streamlining content, and enhancing organization would make the supplementary materials more accessible and valuable for reviewers and readers.

Remarks	Original	Improved	Explanation
Figures lack comprehensive legends and labels, reducing clarity.	Figures 1-4 contain complex diagrams with minimal labels.	Add comprehensive figure legends and labels explaining each component, including color codes, arrows, and modules, to clarify the flow and purpose of each diagram.	Detailed legends and labels improve interpretability and accessibility of visual aids.
Some experimental details are missing, limiting reproducibility.	The supplementary material includes detailed descriptions of the datasets and hyperparameters.	Summarize key dataset details and hyperparameters in a dedicated subsection or table, highlighting the most critical parameters used in experiments, with references to the main text for context.	A clear summary of experimental settings enhances reproducibility and organization.
The supplementary content is verbose, with background sections distracting from core contributions.	The literature review and related work are extensive and somewhat verbose.	Condense the related work to focus on the most directly relevant methods, explicitly contrasting them with the proposed IRON, and move detailed references to an appendix if necessary.	Streamlining background content keeps the focus on supplementary contributions and improves coherence.

R1 - Originality and Contribution

The manuscript presents a notable advancement in weakly-supervised video grounding by integrating iterative proposal refinement with explicit semantic and conceptual correspondence modeling. The approach, while building on established techniques such as leveraging pre-trained vision-language models and iterative refinement, distinguishes itself through its dual-branch distillation and modular design. However, the novelty claim could be further strengthened by more clearly differentiating IRON from similar recent works and explicitly articulating the unique technical innovations. The paper is well-positioned in the literature, but would benefit from a more detailed comparative analysis and a clearer explanation of how its contributions advance the state of the art.

Remarks	Original	Improved	Explanation
Clarify the specific innovation and how it differs from existing pipelines, emphasizing the novelty.	To date, most of the state-of-the-art WSVG methods follow a two-stage pipeline, i.e., firstly generating potential temporal proposals and then grounding with these proposal candidates.	While many recent WSVG methods adopt a two-stage pipeline, our approach uniquely integrates iterative proposal refinement with explicit semantic and conceptual correspondence modeling, setting it apart from prior methods that rely solely on static proposal generation.	This revision clarifies the specific innovation and highlights how the proposed approach differs from existing methods, strengthening the novelty claim.
Highlight the dual-branch distillation as a unique technical contribution.	The proposed approach sets up two lightweight distillation branches to uncover the cross-modal correspondence on both the semantic and conceptual levels.	Our approach uniquely employs dual lightweight distillation branches that explicitly model cross-modal correspondence at semantic and conceptual levels, providing a more in-depth and interpretable alignment than previous methods that often rely on implicit or high-level similarity measures.	This makes the technical innovation more explicit and positions the dual-branch design as a key contribution.
Strengthen the comparative positioning within the literature.	The literature review discusses existing proposal generation and VL transfer methods but lacks a detailed comparison highlighting how IRON's iterative re finement and distillation strategies outperform or differ from similar recent approaches, especially in terms of explicit correspondence modeling.	Include a dedicated subsection contrasting IRON's dual correspondence modeling and iterative refinement with recent VL-guided localization and proposal refinement methods, clearly articulating the unique mechanisms and performance improvements.	A more explicit comparative analysis will clarify the manuscript's originality and contribution relative to the state of the art.

R2 - Impact and Significance

The manuscript offers a substantial contribution to the field of weakly-supervised video grounding by proposing a robust, iterative proposal refinement framework that leverages cross-modal knowledge distillation and label propagation. The technical innovations are poised to influence future research directions and practical applications, such as video retrieval and content moderation. However, the broader societal, industry, and policy implications are not thoroughly articulated, and the discussion could better highlight how the work advances theoretical understanding and opens new research avenues. Explicitly addressing these aspects would further elevate the significance and adoption potential of the research.

Remarks	Original	Improved	Explanation
Explicitly discuss broader societal and industry implications.	The introduction emphasizes the importance of WSVG but does not elaborate on societal impacts.	Add a paragraph discussing how improved video grounding can enhance applications like video retrieval, content moderation, and assistive technologies, thereby emphasizing societal benefits.	Highlighting broader implications increases the perceived significance and relevance of the work.
Clarify the impact on future research directions.	The literature review mentions existing drawbacks but lacks discussion on future research directions.	Incorporate a discussion on how IRON's framework can inspire future research into multimodal pre-training, real-time grounding, or cross-domain adaptation.	Guiding future research efforts positions the work as a foundation for subsequent innovations.
Strengthen the impact statement in the abstract.	The abstract could better emphasize the broader impact of the proposed method.	Revise to explicitly state how IRON's improved proposal coverage and correspondence modeling can transform video understanding and retrieval applications at scale.	A stronger impact statement in the abstract will highlight the work's significance for both industry and academia.

R3 - Ethics and Compliance

While the manuscript demonstrates strong technical innovation and thorough experimentation, it falls short in explicitly addressing key ethical considerations. There is no clear disclosure of conflicts of interest, data privacy measures, or informed consent regarding the use of human video data. Additionally, the manuscript lacks statements on adherence to ethical guidelines or IRB approval. Addressing these issues through clear disclosures and procedural details would elevate the research's ethical standards and compliance, ensuring transparency and trustworthiness.

Remarks	Original	Improved	Explanation
Explicitly address data privacy and consent procedures.	The datasets in use are Charades-STA and ActivityNet Captions.	The datasets used are publicly available, and the authors confirm that data collection adhered to applicable ethical guidelines, including privacy protections and consent procedures where applicable.	Explicit dataset compliance statements reassure readers of ethical adherence regarding human data.
Disclose conflicts of interest and funding sources.	No explicit conflicts of interest are disclosed.	The authors declare that there are no conflicts of interest or funding sources that could have influenced the research outcomes.	Explicit conflict of interest disclosure aligns with best ethical practices and enhances transparency.
Include a formal ethical compliance statement.	The paper does not specify IRB approval or ethical review.	The authors affirm that all data collection and usage procedures comply with institutional review board (IRB) or equivalent ethical standards, and have obtained necessary approvals or waivers.	A formal ethical statement ensures transparency about ethical oversight and adherence to research guidelines.

R4 - Data and Code Availability

The manuscript makes a reasonable effort toward data and code sharing by promising open-source code. However, it lacks detailed descriptions of dataset access, licensing, environment setup, and comprehensive documentation. There is no explicit information on dataset accessibility, usage instructions, or licensing, which limits reproducibility and transparency. Addressing these issues by providing explicit links, licensing information, environment configurations, and detailed instructions would significantly improve the openness and reproducibility of the research.

Remarks	Original	Improved	Explanation
Clarify dataset accessibility and provide access instructions.	The datasets are used for training and evaluation.	The datasets used (Charades-STA and ActivityNet Captions) are publicly available. Dataset access instructions, download links, and licensing conditions will be provided in the supplementary materials or the project repository.	Explicit dataset accessibility and instructions improve transparency and facilitate independent validation.
Enhance code documentation and reproducibility.	The code will be available at https://github.com/mengcaopku/IRON.	The code, including training scripts, pre-processing tools, and environment setup instructions, will be publicly released on GitHub at https://github.com/mengcaopku/IRON, with version tags and detailed README documentation.	Comprehensive code release details and documentation ensure that other researchers can easily reproduce the experiments.
Clarify licensing and restrictions for data and code.	No mention of licensing or access restrictions.	All datasets and code will be released under open licenses (e.g., MIT, CC BY), with explicit statements on any restrictions, ensuring clarity for users.	Clear licensing information supports legal and ethical use, fostering open science.

R5 - Statistical Rigor

The manuscript demonstrates strong methodological innovations and empirical results, but lacks explicit verification of statistical assumptions, justification for hyperparameters, and formal significance testing. There is no discussion of statistical power, effect sizes, or corrections for multiple comparisons. Incorporating these statistical rigor practices would elevate the scientific validity and reproducibility of the work, moving it toward an excellent rating. Currently, the statistical rigor is average, with room for improvement in transparency and robustness.

Remarks	Original	Improved	Explanation
Verify assumptions underlying similarity measures and distillation targets.	The encoded video feature is represented as v RT C, where T is the number of sampled frames and C is the feature dimension.	Explicitly verify the assumptions underlying the similarity measures, such as normality or independence, by conducting normality tests or analyzing the distribution of similarity scores. Incorporate statistical tests or diagnostics to validate the appropriateness of the similarity metrics used for distillation.	This ensures the validity of the similarity-based distillation targets and enhances the robustness of the proposal refinement process.
Justify sample size and hyperparameter choices.	We set the proposal number N to 8 for each video in both datasets.	Provide a sample size justification for the choice of N=8 proposals, possibly through a power analysis or ablation studies demonstrating the impact of different proposal counts on performance and stability.	Clarifies the rationale behind hyperparameter selection and supports the statistical robustness of the method.
Apply statistical significance testing and correction for multiple comparisons.	Extensive experiments and ablation studies on two challenging WSVG datasets have attested to the effectiveness of our IRON.	Apply statistical significance testing (e.g., paired t-tests, bootstrap confidence intervals) to compare the proposed method against baselines or ablations, correcting for multiple comparisons where applicable.	Provides rigorous statistical validation of the reported improvements, reducing the risk of false positives.

R6 - Technical Accuracy

The manuscript demonstrates a high level of technical rigor, with innovative ideas and comprehensive experiments. The methodology is sound and results are convincing. However, minor clarifications are needed in mathematical derivations, algorithm details, and the discussion of limitations. Explicitly specifying matrix dimensions, proposal formats, and providing deeper analysis of ablation results would further improve reproducibility and technical clarity. Addressing these points would elevate the work to an excellent standard.

Remarks	Original	Improved	Explanation
Clarify mathematical operations and matrix dimensions in equations.	Equations (1)-(3) combine sigmoid activations with linear transformations but lack detailed explanation of the transformations.	Provide explicit definitions of the linear transformations, including whether biases are used, and clarify the role of the sigmoid activation in bounding the scores between 0 and 1. For example: 'The linear transformations Wk_s, Wk_e, Wk_c are matrices of size C×1 (or C×M for Wk_c) with optional bias terms, followed by sigmoid activation to produce scores in [0,1].'	Clarifies the mathematical operations, aiding reproducibility and understanding of the distillation process.
Specify proposal format and IoU calculation in the label propagation algorithm.	Algorithm 1 does not specify how IoU is calculated or the format of proposals.	Add detailed comments or steps in the pseudo-code explaining that proposals are represented as coordinate pairs (start, end), and IoU is computed as the intersection over union of these intervals. For example: 'Calculate IoU between proposals u_n and u_ik as the ratio of the intersection length over the union length of their temporal intervals.'	Ensures correct implementation and reduces ambiguity in the label propagation process.
Define key technical terms clearly.	Terms like 'semantic distillation' and 'proposal coverage' are used but not explicitly defined.	Include precise definitions early in the methodology: 'Semantic distillation refers to aligning proposal features with query semantics via similarity scores estimated by the VL model. Proposal coverage indicates the extent to which a proposal overlaps with the ground truth or relevant event segments, measured by IoU.'	Clarifies key concepts for better understanding and precise communication.

R7 - Consistency

The manuscript demonstrates strong logical coherence and consistency across sections, with clear methodology and compelling experimental validation. Minor improvements are needed in explicitly linking methods to results, providing clearer definitions, and balancing the discussion of limitations. The figures, tables, and citations are generally well-aligned and support the narrative effectively. Addressing the suggested refinements would further enhance the clarity, transparency, and scholarly rigor of the presentation.

Remarks	Original	Improved	Explanation
Explicitly link ablation results to the iterative process.	While the methodology describes the iterative re finement process, the results section (4.2, Table 2a) shows ablation results that suggest the importance of different loss components, but lacks explicit connection to the specific iterative steps or how each iteration impacts the proposals' quality.	Explicitly link the ablation results to the iterative process by discussing how each loss component influences proposal refinement over iterations, e.g., 'Removing semantic loss reduces the proposal coverage in subsequent iterations, as shown by the decline in R1@0.3.'	This clarifies the causal relationship between methodology components and observed outcomes, strengthening the methods-results coherence.
Balance conclusions with acknowledgment of limitations.	The conclusions about the superiority of IRON are based on quantitative metrics, but the discussion does not sufficiently address potential limitations or failure cases.	Add a paragraph acknowledging limitations such as cases where proposals might still be biased or coverage incomplete, and suggest future directions to address these issues.	This balances the results and strengthens the credibility of conclusions by providing a nuanced perspective.
Improve logical flow from problem statement to solution.	The transition from discussing the drawbacks of existing methods to introducing IRON could be smoother, especially in explicitly linking how each component of IRON addresses specific issues.	Include a summarizing paragraph that explicitly states: 'To address lack of explicit correspondence modeling, IRON employs semantic and conceptual distillation; to mitigate partial coverage, it uses iterative label propagation...'	This enhances logical flow by clearly mapping problems to solutions, aiding reader comprehension.

W1 - Language and Style

The manuscript demonstrates a strong command of technical language and generally adheres to academic writing standards. Positive aspects include clear technical descriptions and mostly correct grammar. However, there are minor issues with punctuation, sentence complexity, and occasional informality or inconsistent capitalization. Improving sentence clarity, punctuation consistency, and formal tone would further elevate the manuscript's quality and readability.

Remarks	Original	Improved	Explanation
Redundant and informal adverb usage in formal writing.	i.e., firstly generating potential temporal proposals	i.e., generating potential temporal proposals	Removing 'firstly' streamlines the phrase and aligns with formal academic style, avoiding redundancy.
Overly complex sentence structure reduces readability.	The encoded video feature is represented as v ∈RT ×C, where T is the number of sampled frames3 and C is the feature dimension.	The encoded video feature is represented as v ∈ ■^{T×C}, where T is the number of sampled frames, and C is the feature dimension.	Splitting into two sentences or adding commas improves readability; also, using proper notation for sets (e.g., ■^{T×C}) enhances clarity in technical writing.
Pronoun ambiguity and lack of subject clarity.	Our label propagation strategy is based on two assumptions of the proposal with minimum distillation loss: 1) High hit rate. They can always hit the region of interest.	Our label propagation strategy is based on two assumptions regarding proposals with minimum distillation loss: 1) a high hit rate, meaning they can reliably identify the region of interest.	Clarifies the subject ('proposals') and improves grammatical agreement; also, enhances formal tone and clarity of the assumption.

W2 - Narrative and Structure

The manuscript presents a comprehensive and technically detailed approach but would benefit from improved narrative coherence and logical flow. The main issues are abrupt transitions between sections, insufficient integration of evidence with claims, and a lack of explicit links between hypotheses and experimental validation. Positive aspects include a logical overall structure and thorough experimental coverage. Enhancing transitions, clarifying the purpose of each section, and explicitly connecting results to research questions would strengthen the narrative and make the paper more compelling.

Remarks	Original	Improved	Explanation
Abrupt transitions between methodology subsections impede narrative flow.	Section 3 (Approach) is dense with technical details but lacks transition sentences between subsections.	After describing the proposal generation process, we now introduce the semantic and conceptual distillation targets, which serve as the foundation for our iterative refinement. Subsequently, we detail the label propagation strategy that refines proposal confidence scores across iterations.	Adds connective tissue that guides the reader through the methodological steps, improving flow.
Experimental results are presented without sufficient narrative commentary linking them to claims.	Results are densely presented with tables but lack narrative commentary.	Our experimental results demonstrate that IRON consistently outperforms previous methods across multiple metrics. For instance, on Charades-STA, our method achieves an R1@0.3 of 70.71%, surpassing the previous best CPL by over 4%. These results validate the effectiveness of our iterative proposal refinement strategy.	Explicitly links data to claims, enhancing clarity and persuasive power.
Conclusion does not explicitly revisit research questions or hypotheses.	The conclusion summarizes contributions but does not revisit initial research questions.	In conclusion, our IRON effectively addresses the challenges of proposal coverage and correspondence modeling in weakly-supervised video grounding, as demonstrated by extensive experiments. Future work will explore integrating more sophisticated language understanding modules to further improve localization accuracy.	Revisits research aims and ties them to results, closing the narrative loop.

W3 - Clarity and Conciseness

The manuscript contains strong technical content but would benefit from greater clarity and conciseness. The use of complex language, unexplained jargon, and long sentences or paragraphs can hinder accessibility, especially for non-expert readers. Positive aspects include precise technical descriptions and comprehensive coverage. Simplifying language, breaking up dense paragraphs, and providing definitions for technical terms would improve readability and make the work more accessible to a broader audience.

Remarks	Original	Improved	Explanation
Complex and technical phrases reduce accessibility.	Weakly-Supervised Video Grounding (WSVG) aims to localize events of interest in untrimmed videos with only video-level annotations.	Weakly-Supervised Video Grounding (WSVG) seeks to locate specific events in untrimmed videos using only video-level labels.	Simplifies language for clarity and reduces technical complexity while maintaining meaning.
Heavy reliance on unexplained jargon.	Some technical terms like 'IoU,' 'distillation loss,' and 'proposal features' are used frequently without immediate explanation.	Introduce brief definitions or explanations of technical terms like 'IoU' (Intersection over Union) and 'distillation loss' when first mentioned.	Clarifies jargon for readers unfamiliar with specific terminology, improving accessibility.
Long, dense paragraphs hinder readability.	Many paragraphs, especially in the 'Related Work' and 'Approach' sections, are very long and dense, making it difficult for readers to identify main ideas.	Break long paragraphs into shorter ones, each focusing on a single main idea or aspect, to enhance readability and comprehension.	Shorter paragraphs improve visual clarity and help readers follow complex arguments.

W4 - Terminology Consistency

The manuscript demonstrates solid technical content but suffers from inconsistent terminology, notation, and abbreviation usage. Inconsistent use of terms like 'proposal' and 'candidate proposal,' variable naming, and mathematical notation can confuse readers. Positive aspects include generally correct field-specific terminology. Standardizing terminology, notation, and acronym definitions throughout the paper will improve clarity and professional presentation.

Remarks	Original	Improved	Explanation
Inconsistent use of the terms 'proposal' and 'candidate proposal' creates ambiguity.	Proposal candidates	Proposal (candidate proposals)	Adding parentheses clarifies that 'candidate proposals' is an explanatory phrase, improving readability and consistency.
Mathematical notation for vectors and matrices is inconsistent.	p	bold p (e.g., **p**)	Using boldface for vectors/matrices consistently (e.g., **p**) enhances notation clarity throughout equations.
Acronyms are sometimes used before being defined.	WSVG	Weakly-Supervised Video Grounding (WSVG)	Always defining acronyms at first use ensures clarity, especially when used multiple times, and helps readers unfamiliar with the term.

W5 - Inclusive Language

The manuscript uses neutral, technical language and avoids overtly exclusive or biased terms. However, it could be more inclusive by explicitly acknowledging cultural, linguistic, and accessibility limitations, especially regarding dataset diversity and language choices. Positive aspects include the absence of gendered language and the use of open-source practices. Explicitly addressing inclusivity in datasets, language, and accessibility would broaden the work's applicability and demonstrate a commitment to equitable research.

Remarks	Original	Improved	Explanation
Lack of acknowledgment of linguistic limitations and inclusivity.	We chose DistilBERT [46] pre-trained on English Wikipedia and Toronto Book Corpus for its lightweight model capacity.	We selected DistilBERT, pre-trained on English Wikipedia and the Toronto Book Corpus, recognizing that this focus on English-language datasets may limit linguistic diversity; future work could incorporate multilingual models to enhance inclusivity.	Acknowledging language limitations and suggesting inclusivity through multilingual approaches broadens the scope and accessibility of the research.
Datasets are predominantly Western and English-centric.	The proposed IRON model is evaluated on datasets primarily consisting of Western content and English language queries.	The evaluation of IRON is conducted on datasets predominantly featuring Western content and English language queries; expanding to include datasets from diverse cultural backgrounds and languages would improve global applicability.	Explicitly recognizing cultural and linguistic limitations encourages future inclusivity and broader relevance.
No explicit mention of accessibility for researchers with disabilities.	Throughout the paper, technical terminology is used without explicit consideration of accessibility for non-expert audiences.	Throughout the paper, efforts will be made to clarify technical terminology and include accessible explanations to support understanding among diverse audiences, including researchers from different backgrounds and with varying levels of expertise.	Enhancing clarity and accessibility promotes inclusivity across diverse reader groups.

W6 - Citation Formatting

The manuscript generally follows numeric citation conventions but exhibits inconsistencies in the use of brackets, placement, and the handling of multiple references. There are also occasional inconsistencies in the use of 'et al.' and cross-referencing. Positive aspects include comprehensive referencing of related work. Standardizing citation formatting, ensuring consistent placement, and cross-verifying all references will improve professionalism and scholarly integrity.

Remarks	Original	Improved	Explanation
Inconsistent use of brackets and parentheses for multiple citations.	[21, 37, 41, 72, 73]	(e.g., [21, 37, 41, 72, 73])	Standardizes citation delimiters, ensuring all multiple citations are consistently enclosed in brackets, enhancing clarity.
Inconsistent placement of citations within sentences.	video retrieval [13, 19]	video retrieval (e.g., [13, 19])	Aligns with the style of citing multiple references within parentheses, maintaining consistency across the manuscript.
Potential mismatches or missing references due to lack of cross-verification.	Some in-text citations (e.g., '[37]') are not cross-verified with the reference list, risking mismatches or missing references.	Implement a systematic cross-checking process or use automated reference management to ensure all citations are accurate and correspond to the correct entries.	Ensures scholarly rigor and prevents referencing errors.

W7 - Target Audience Alignment

The manuscript is well-aligned with an expert audience in video-language understanding, such as the CVPR community. It demonstrates technical rigor and comprehensive experimental validation. However, some methodology sections assume advanced background knowledge, and figures could be more self-explanatory. Including more background explanations and enhancing figure captions would improve accessibility for a broader scholarly audience while maintaining depth for experts.

Remarks	Original	Improved	Explanation
Assumes advanced background knowledge in methodology sections.	The methodology description assumes familiarity with advanced concepts.	Include brief background explanations or references for complex concepts such as cross-modal similarity estimation, transformer proposal generation, and distillation strategies, to make the methodology accessible to a broader audience.	Expanding background details ensures that readers from adjacent fields or early-stage researchers can follow the technical content more easily.
Figures lack comprehensive captions or explanations.	The figures are referenced but lack comprehensive captions or explanations.	Enhance all figures with detailed captions that explicitly describe what each diagram illustrates, including the key components and their roles, to facilitate independent understanding without extensive cross-referencing.	Improves visual comprehension and makes visuals more self-contained, benefiting diverse learning styles.
Reference list could include more recent or seminal works on emerging topics.	The reference list includes many relevant works but lacks recent or seminal papers on emerging topics like open-vocabulary detection and recent transformer architectures.	Expand the reference list to include recent and seminal works on open-vocabulary detection and transformer-based models to enhance contextual depth and demonstrate engagement with the latest research trends.	Including up-to-date references positions the work within current research and increases its relevance.