Weakly-Supervised Spoken Video Grounding via Semantic Interaction Learning

Ye Wang[*], **Wang Lin**[*], **Shengyu Zhang**[*],
Tao Jin, **Linjun Li**, **Xize Cheng**, **Zhou Zhao**[†]
Zhejiang University
{yew,linwanglw,sy_zhang}@zju.edu.cn
{jint_zju,lilinjun21,chengxize,zhaozhou}@zju.edu.cn

Abstract

The task of spoken video grounding aims to localize moments in videos that are relevant to descriptive spoken queries. However, extracting semantic information from speech and modeling the cross-modal correlation pose two critical challenges. Previous studies solve them by representing spoken queries based on the matched video frames, which require tremendous effort for frame-level labeling. In this work, we investigate weakly-supervised spoken video grounding, i.e., learning to localize moments without expensive temporal annotations. To effectively represent the cross-modal semantics, we propose Semantic Interaction Learning (SIL), a novel framework consisting of the acoustic-semantic pre-training (ASP) and acoustic-visual contrastive learning (AVCL). In ASP, we pre-train an effective encoder for the grounding task with three comprehensive tasks, where the robustness task enhances stability by explicitly capturing the invariance between time- and frequency-domain features, the conciseness task avoids over-smooth attention by compressing long sequence into segments, and the semantic task improves spoken language understanding by modeling the precise semantics. In AVCL, we mine pseudo labels with discriminative sampling strategies and directly strengthen the interaction between speech and video by maximizing their mutual information. Extensive experiments demonstrate the effectiveness and superiority of our method.[1]

1 Introduction

Temporal video grounding (Gao et al., 2017; Hendricks et al., 2017) is an important task in the cross-modal understanding field (Zhang et al., 2020d; Jin et al., 2020; Xun et al., 2021; Jin and Zhao, 2021; Yin et al., 2022), aiming to retrieve a target moment within a video based on a given query. With the progress in deep learning, there has been significant achievements in this area. While most previous studies focus on textual queries, recent work (Xia et al., 2022) introduce the spoken video grounding task by incorporating the spoken query into the video grounding, as shown in Figure 1.

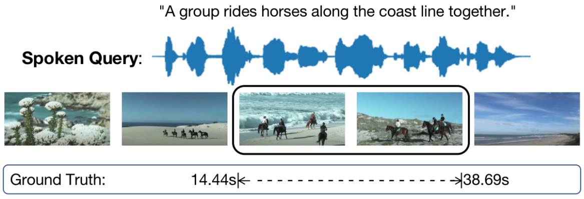


Figure 1: An Example of Spoken Video Grounding.

However, such video grounding task with spoken queries presents unique challenges compared to its text-based counterpart. First, the encoding of speech is inherently difficult due to its weak and volatile semantic information, making it arduous to extract useful features for video grounding. Second, even with the acquisition of valuable semantic features, modeling speech-video interaction and extracting cross-modal content still poses an inevitable obstacle.

Prior work (Xia et al., 2022) address these two problems simultaneously through the proposed video-guided contrastive predictive coding, which utilizes aligned video clips to learn semantic representations from spoken queries. However, a critical drawback is its heavy reliance on precise temporal matching between video and spoken queries. The acquisition of such fine-grained annotations requires substantial manual labor, hindering the practicality and applicability of this approach.

In this work, we address the issue of intensive labor by investigating a novel task called Weakly-Supervised Spoken Video Grounding (WSVG), aiming to localize speech-aligned moments in videos under a weakly-supervised scheme. In this setting, we only have access to aligned video-

[*] Equal contribution.
[†] Corresponding author
[1] https://github.com/yewzz/SIL.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 10914–10932
July 9-14, 2023 ©2023 Association for Computational Linguistics

speech pairs during training without any temporal annotations, which poses a greater challenge.

To tackle the aforementioned problem, we propose a novel framework Semantic Interaction Learning (SIL), following a progressive pipeline to first pre-train a speech encoder for semantic encoding and then learn speech-video alignment during grounding training. It consists of two key components: acoustic-semantic pre-training (ASP) and acoustic-visual contrastive learning (AVCL).

In the pre-training stage, the ASP module utilizes external speech-text data to train a speech encoder capable of extracting rich semantic information from speech. To adapt the encoded features for the downstream weakly-supervised video grounding, ASP includes three specialized tasks targeting three specific characteristics. (1) The **robustness task** focuses on the encoder's ability to handle complex and noisy speech, which is a practical problem in the real world. Considering time series data can be split into the time and frequency domains to provide invariance regardless of varying distributions (Zhang et al., 2022), we utilize both time- and frequency-based speech feature for pre-training and forces their encoded semantic information to be consistent. (2) The **conciseness task** addresses the issue of long sequence features with redundant information, which results in scattered distribution of the video-speech attention and impedes effective interaction. Hence, we compress the encoded features into discrete segments via I&F algorithm (Dong and Xu, 2020), refining the feature sequence for effective interaction. (3) The **semantic task** emphasizes the extraction of key semantics for grounding, which is a crucial requirement in this fine-grained cross-modal understanding task. Unlike the trivial self-supervised method (Baevski et al., 2020) or knowledge distillation method (Hinton et al., 2015), we draw inspiration from the human understanding system that encompasses auditory perception and cognitive processing (Dong et al., 2021). Concretely, we introduce a connectionist temporal classification (CTC) (Graves et al., 2006) loss to facilitate training, and further consider both sequence-level and word-level semantic to ensure the comprehensive semantic transfer.

In the grounding stage, the AVCL module directly enhances the correlation between video and speech. Despite the effective semantic encoding of spoken queries, the discrepancy between video and speech still hinders the cross-modal interaction. As video and speech are from two distinct feature spaces, AVCL leverages contrastive learning to maximize their agreement. First, we perform location-based selection and score-based mining to select pseudo labels with high confidence. With the located boundary of the predicted pseudo proposal, we can coarsely select the negative samples from regions outside and further calculate the clip-level score inside the boundary to mine positive/negative samples. Then, based on these discriminative samples, we contrastively maximize the mutual information between speech and positive clips.

Our main contributions are listed as follows:

- We investigate a new task WSVG to explore the weakly-supervised spoken video grounding.
- We propose a novel framework SIL to effectively model the semantic contents of video-speech interaction, where the ASP module enhances semantic encoding and the AVCL module improves cross-modal interaction.
- Extensive experiments verify the superiority of our approach in terms of both accuracy and efficiency.

2 Related Works

2.1 Temporal Video Grounding

Temporal video grounding aims to localize the moment corresponding to the query. Under the supervised setting, existing methods can be categorized into the top-down and bottom-up frameworks. The top-down methods (Gao et al., 2017; Hendricks et al., 2017; Liu et al., 2018; Chen et al., 2018; Zhang et al., 2019) first generate proposals and then estimate cross-modal alignment scores for them. And the bottom-up methods (Chen et al., 2019a, 2020; Wu et al., 2020; Zhang et al., 2020a; Zhao et al., 2021) directly calculate the frame-level probabilities of being temporal boundaries. Under the weakly-supervised setting, the methods can be categorized into the multiple instance learning (MIL) and the reconstruction frameworks. The MIL methods learn the latent visual-textual alignment by distinguishing the matched video-language pairs from the unmatched pairs. For example, Gao et al. 2019 devise an alignment and a detection module. Zhang et al. 2020d develop contrastive learning between counterfactual results. Huang et al. 2021 explore cross-sentence relational constraints. The reconstruction methods reconstruct the query from visual contents during training and utilize intermediate results to localize. For example, Lin et al. 2020 utilize language reconstruction to rank proposals. Song et al. 2020 further employ the attention weight. Zheng et al. 2022 mine negatives within the same video. Recent work (Xia et al., 2022) study the spoken video grounding task and represent speech with video-guided contrastive predictive coding. We consider the intensive labor and introduce the weakly-supervised setting.

2.2 Vision-Audio Learning

Vision-audio learning has attracted researchers' interest in recent years. Since Harwath and Glass 2015 collect spoken captions for Flickr8k, much research (Chrupała, 2022; Harwath et al., 2019; Higy et al., 2021; Scholten et al., 2021) begins to attach importance to this field. Some works emphasize the cognitive and linguistic questions, such as understanding how different learned layers correspond to visual stimuli (Chrupała et al., 2017; Gelderloos and Chrupała, 2016), learning linguistic units (Harwath and Glass, 2019; Harwath et al., 2019). Oncescu et al. 2020 propose QuerYD, a video dataset with both text and audio descriptions for text-video retrieval and corpus moment retrieval. Recent work (Cheng et al., 2023) study visual speech translation and recognition.

3 Methods

3.1 Overview

Problem Formulation. Given an untrimmed video V and a spoken query S, this task aims to train a network $G(V, S)$ to localize the most relevant moment proposal p corresponding to the spoken query S, without (p, S) alignment annotations, i.e. only the video-speech pair (V, S) are available.

Overall Pipeline. Our SIL follows a two-stage pipeline. First, we pre-train the speech encoder with external speech-text data. In this stage, the ASP module develops three tasks to improve robustness, conciseness and semantic respectively, which enables the encoder to extract effective information of speech for the downstream task. Then, we fix the speech encoder and conduct weakly-supervised training on the grounding dataset via our base network. In this stage, the ACVL module selects contrastive samples via a discriminative sampling strategy and then maximizes the mutual information between video and speech.

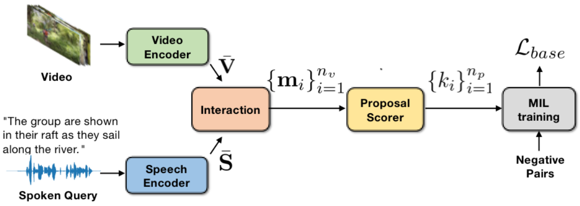


Figure 2: The Base Network for WSVG.

3.2 Base Network

To illustrate our framework clearly, we first formulate the base grounding network $G(V, S)$ under SIL as following four common modules:

- Feature Encoder: The video encoder encodes video features as $\bar{\mathbf{V}} = \{\bar{\mathbf{v}}_i\}_{i=1}^{n_v} \in \mathbb{R}^{n_v \times d}$ and the speech encoder encodes speech features as $\bar{\mathbf{S}} = \{\bar{\mathbf{s}}_i\}_{i=1}^{n_c} \in \mathbb{R}^{n_c \times d}$, where d is the hidden size, n_v and n_c are the length of video and speech features, respectively.
- Interaction: It develops the cross-modal interaction between $\bar{\mathbf{V}}$ and $\bar{\mathbf{S}}$, then outputs multi-modal clip features $\{\mathbf{m_i}\}_{i=1}^{n_v} \in \mathbb{R}^{n_v \times d}$. The interaction methods include attention-based aggregation and feature fusion (Zhang et al., 2019).
- Proposal Scorer: Based on multi-modal clip features $\{\mathbf{m_i}\}_{i=1}^{n_v}$, it extracts n_p proposal features and calculates their alignment scores $K = \{k_i\}_{i=1}^{n_p}$. The score of each video-speech pair (V, S) is $f(K)$, where $f(\cdot)$ is the average of the top-R proposal scores.
- Training: We follow the MIL paradigm (Zhang et al., 2020c) to utilize the score $f(K)$ to train the model with binary cross entropy loss $\mathcal{L}_{\text{base}}$, which distinguishes the matched video-speech pair (V, S) from two randomly-selected negative pairs (V', S) and (V, S').

The details are introduced in Appendix A.

3.3 Acoustic-semantic Pre-training

In this section, we elaborate on our pre-training for the speech encoder. Given the external data, we denote the speech as S and its paired text (i.e. transcript) as W. We first introduce the overall encoding process and then detail our designed tasks.

Our speech encoder consists of convolutional layers and N_a+N_s layers Transformer encoder (Vaswani et al., 2017). (1) First, the robustness task in Section 3.3.1 simultaneously considers the time-based features S^{time} and frequency-based features S^{freq} as the speech input. For ease of presentation, we omit superscripts and denote them

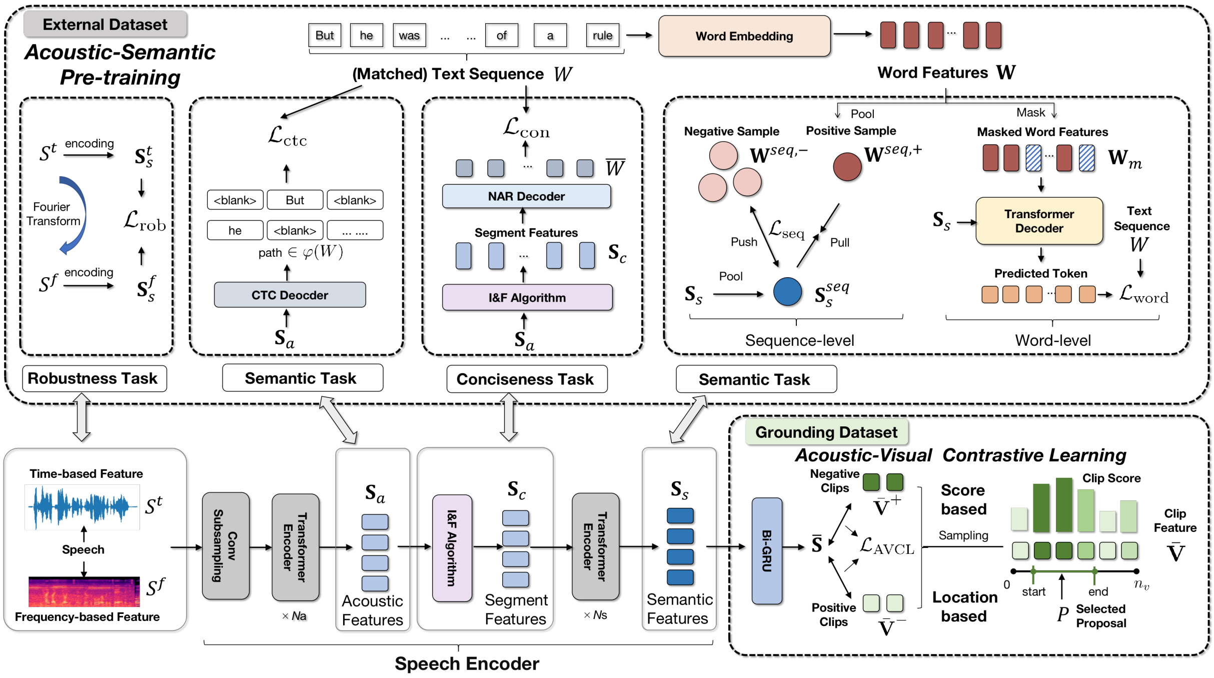


Figure 3: The Concrete Design of Semantic Interaction Learning Framework.

as S. We apply convolutional blocks on S to extract their deep features $\mathbf{S} = \{\mathbf{s}_i\}_{i=1}^{n_s} \in \mathbb{R}^{n_s \times d}$, where n_s is the downsampled sequence length. (2) Next, we input these features into N_a layers Transformer encoder to obtain acoustic features $\mathbf{S}_a \in \mathbb{R}^{n_s \times d}$. The conciseness task in Section 3.3.2 then compresses them into segment-level features $\mathbf{S}_c \in \mathbb{R}^{n_c \times d}$, where n_c is the length of segments. (3) Finally, we input $\mathbf{S}_c$ into another N_s layers Transformer encoder to learn semantic features $\mathbf{S}_s \in \mathbb{R}^{n_c \times d}$. We train $\mathbf{S}_a$ and $\mathbf{S}_s$ via the semantic task in Section 3.3.3.

3.3.1 Robustness Task

As explicit consideration of the frequency domain provides an understanding of the behavior of time series that cannot be captured in the time domain (Zhang et al., 2022), we aim to improve robustness by considering both domain features and identifying their general property of speech that is preserved across transformations.

For each speech S, we generate its time-based feature as S^t (i.e. wave) and frequency-based feature as S^f (i.e. mel-frequency spectrum). They can be converted to each other through Fourier transform and inverse Fourier transform. Then we simultaneously input two features into the speech encoder and perform the same aforementioned pre-training. Here we adopt different convolutional layers and N_a layers transformer encoder to model acoustic property for two distinct domain features, while we remain the rest N_s layers identical for semantic sharing. Following the encoding process, we can obtain their corresponding semantic features $\mathbf{S}_s^t$ and $\mathbf{S}_s^f$, which are output from the last N_s layers encoder. To learn the invariance across domains, we apply L1 loss to align two features in a common feature space, given by:

$$\mathcal{L}_{\text{rob}} = |\mathbf{S}_s^t - \mathbf{S}_s^f| \tag{1}$$

After pre-training, we yield the final semantic features $\tilde{\mathbf{S}}$ via concatenation as $\tilde{\mathbf{S}} = [\mathbf{S}_s^t, \mathbf{S}_s^f]$. During grounding training, we further encode it as $\bar{\mathbf{S}}$ by a Bi-GRU network.

3.3.2 Conciseness Task

The long sequence of speech may result in over-smooth attention distribution (Touvron et al., 2023). To alleviate ineffective cross-modal interaction caused by this, we design a conciseness task to compress long acoustic features into segments.

We adopt continuous integrate-and-Fire (I&F) (Dong and Xu, 2020) algorithm, which is a soft and monotonic alignment mechanism. First, the input features $\mathbf{S}_a = \{\mathbf{s}_{a,i}\}_{i=1}^{n_s}$ are fed to a weight predictor to obtain the weights $G = \{g_i\}_{i=1}^{n_s}$, representing the amount of information in $\mathbf{S}_a$. We then scan and accumulate them from left to right until the sum reaches the threshold(set to 1.0), indicating a semantic boundary b_j is detected. Then, we reset the accumulation and continue to scan the rest which begins with r_j. Finally, we multiply all g_i by corresponding $\mathbf{s}_{a,i}$ and integrate them to obtain segment features $\mathbf{S}_c = \{\mathbf{s}_{c,i}\}_{i=1}^{n_c}$, where n_c is the detected segment number.

To enable each segment carry the complete semantic information, we regard each word in the text sequence $W = \{w_i\}_{i=1}^{n_w}$ as an independent supervision signal. Then we develop a non-auto-regressive decoder to predict word tokens $\bar{W} = \{\bar{w}_i\}_{i=1}^{n_{\bar{w}}}$ from the segment features. The alignment loss consists of two terms:

$$\mathcal{L}_{\text{con}} = (n_{\bar{w}} - n_w) - \sum_{(x,y)\in(\bar{W},W)} \log P_{nar}(y|x) \quad (2)$$

where the first item aims to force the length of predicted token consistent with the target text and the second item is the cross entropy loss for word recognition.

3.3.3 Semantic Task

We design the semantic task to transfer the knowledge from the text representation model, e.g. Glove embedding (Pennington et al., 2014), to the encoded speech features $\mathbf{S}_s$. To stabilize and facilitate semantic learning, we first utilize an ordinary CTC loss without considering the syntactic structure and semantic knowledge of target word sequences. Next, with the embedding features $\mathbf{W} = \{\mathbf{w}_i\}_{i=1}^{n_w}$ of text W, we perform semantic learning with sequence-level and item-level objectives, where the sequence-level objective tries to contrastively align matched speech-text features and the word-level objective aims to reconstruct the masked key word based on the speech. The full semantic loss consists of three terms $\mathcal{L}_{\text{sem}} = \mathcal{L}_{\text{ctc}} + \mathcal{L}_{\text{seq}} + \mathcal{L}_{\text{word}}$.

CTC Warm-up. To model the acoustic structure for semantic learning, we build a CTC decoder over the features $\mathbf{S}_a$ and optimize it with a CTC loss. Given the target word sequence $W = \{w_i\}_{i=1}^{n_w}$, CTC introduces a set of intermediate paths $\varphi(W)$, where each path $C \in \varphi(W)$ is composed of words and blanks that can be reduced to the target sequence. The loss is computed by:

$$\mathcal{L}_{\text{ctc}} = -\log \sum_{C\in\varphi(W)} P(C|\mathbf{S}_a) \quad (3)$$

Sequence-level Contrastive Objective. The sequence-level objective employs contrastive learning to bring the speech closer to its corresponding text in the global feature space. First, we apply mean-pooling on word features $\mathbf{W}$ and speech features $\mathbf{S}_s$ to obtain their sequence-level features $\mathbf{W}^{seq}$ and $\mathbf{S}_s^{seq}$. For each sequence-level speech feature $\mathbf{S}_s^{seq}$, we denote the corresponding text feature as $\mathbf{W}^{seq,+}$ and randomly sample B unmatched text features $\mathbf{W}^{seq,-}$. We adopt the Info-NCE loss (Gutmann and Hyvärinen, 2010; Sun et al., 2019) to optimize the alignment by:

$$\mathcal{L}_{\text{seq}} = -\log \frac{e^{\mathbf{S}_s^{seq} \cdot \mathbf{W}^{seq,+}}}{e^{\mathbf{S}_s^{seq} \cdot \mathbf{W}^{seq,+}} + \sum_{i=1}^{B} e^{\mathbf{S}_s^{seq} \cdot \mathbf{W}^{seq,-}}} \quad (4)$$

Word-level Generative Objective. Though the sequence-level objective ensures the global semantic, it fails to capture the information of crucial word for grounding. Thus, we further leverage the speech content to predict the masked words in order to preserve the word-level knowledge.

We mask $x\%$ of the word features $\mathbf{W}$ to generate modified word features $\mathbf{W}_m$ as (Devlin et al., 2018). Then we build a bi-directional Transformer decoder with $\mathbf{W}_m$ as queries and $\mathbf{S}_s$ as keys and values. The output o of the decoder is given by $o = \text{TransformerDecoder}(\mathbf{W}_m, \mathbf{S}_s)$. We employ a linear layer to predict the word distribution $\{\mathbf{e}_i\}_{i=1}^{n_w} \in \mathbb{R}^{n_w \times d_b}$, where d_b is the vocabulary size. Finally, we compute the negative log-likelihood of each word and add them up, given by:

$$\mathcal{L}_{\text{word}} = -\sum_{i=1}^{n_w - 1} \log p(w_{i+1}|\mathbf{e}_i) \quad (5)$$

3.4 Acoustic-visual Contrastive learning

In grounding training, we conduct acoustic-visual contrastive learning (AVCL). To mine visual samples as guidance, we design two discriminative sampling strategies.

Location-based Selection. As no temporal annotations are provided under the weakly-supervised setting, we consider the selected proposal p as the latent visual guidance and coarsely select negative samples outside the boundary of the proposal p.

Score-based Mining. To mine high-quality visual samples, we further calculate clip-level scores $\{c_i\}_{i=1}^{n_v}$ for clips inside the boundary through the proposal scorer in Section 3.2, where the proposal features is replaced with the clip features as input. Then we select several clips with the highest scores as positive samples, while reserving a subset with the lowest scores as negative samples.

With the above strategies, we select T positive and T negative clips samples. The inspiration comes from the observation on experiments. During early training, the predicted boundary tends to cover a wide temporal range, thus the location-based selection provides insufficient negative samples and the score-based mining can further select hard negative clips within the predicted proposal as a complementary part. As the training goes on, the predicted boundary will narrow and be more accurate, the location-based selection can select enough negative samples to avoid introducing noise.

Given $\bar{\mathbf{V}}^{+}$ and $\bar{\mathbf{V}}^{-}$ as features of positive and negative clips respectively, we maximize the lower bound of cross-modal mutual information through Jensen-Shannon estimator (Hjelm et al., 2018; Nan et al., 2021) as:

$$\mathcal{L}_{\text{AVCL}} = \mathbb{E}[(\phi(\bar{\mathbf{S}}, \bar{\mathbf{V}}^{-}))] - \mathbb{E}[(\phi(\bar{\mathbf{S}}, \bar{\mathbf{V}}^{+}))] \quad (6)$$

where $\phi(\cdot,\cdot)$ is the MI discriminator.

3.5 Training and Inference

Pre-Training. We combine the losses of three tasks to form the overall loss $\mathcal{L}_{\text{ASP}}$ for acoustic-semantic pre-training by:

$$\mathcal{L}_{\text{ASP}} = \lambda_1 \mathcal{L}_{\text{rob}} + \lambda_2 \mathcal{L}_{\text{conc}} + \mathcal{L}_{\text{sem}} \quad (7)$$

Grounding Training. We fix the speech encoder and perform grounding training with the ACVL module. The full loss $\mathcal{L}_{\text{G}}$ is given by:

$$\mathcal{L}_{\text{G}} = \mathcal{L}_{\text{base}} + \lambda_3 \mathcal{L}_{\text{AVCL}} \quad (8)$$

Inference. During the inference, we directly select the proposal with the highest score as the result.

4 Experiments

4.1 Datasets

We evaluate the weakly-supervised spoken video grounding on the ActivityNet Speech dataset and perform pre-training on the LibriSpeech dataset.

ActivityNet Speech (Xia et al., 2022). It is constructed on the ActivityNet Caption (Caba Heilbron et al., 2015), which contains 19,209 videos annotated with 3.65 textual descriptions on average and marked with timestamps. ActivityNet Speech transforms the textual annotations into speech.

LibriSpeech (Panayotov et al., 2015). It is a collection of approximately 1,000 hours of audiobooks in the LibriVox project (Kearns, 2014).

4.2 Implementation Details

For video features, we follow the previous works (Gao et al., 2017) to extract C3D (Tran et al., 2015) features as input. For speech features, we adopt the raw wave as time-based feature and use Fourier Transform to obtain 80-dimensional log-mel spectrograms as frequence-based feature. In pre-training, we use the pre-trained Glove (Pennington et al., 2014) embeddings as the word features. The full details are listed in Appendix B.2

4.3 Evaluation Metrics

For a fair comparison, we follow previous temporal video grounding works (Gao et al., 2017) to employ the **R@n,IoU=m** as the evaluation metrics, which is the percentage of at least one of the top-n moments with the IoU $>$ m. We also report the **mIoU** value which is the average IoU between the top-1 selected moment and the ground truth.

4.4 Performance Comparison

Baseline. Since no existing strategy can be directly applied to WSVG, we consider baselines under the cascaded and end-to-end (E2E) setting.

- **Cascaded methods**: these methods use textual input recognized by the ASR model (Baevski et al., 2020) as input to the grounding network. We select the following weakly-supervised text-based video grounding methods: **WSLLN** (Gao et al., 2019), **RTBPN** (Zhang et al., 2020c) and **SCN** (Lin et al., 2020).
- **E2E methods**: these methods use speech as direct input to the grounding network. (1) We consider the supervised approach **VSLNet** (Zhang et al., 2020a) and **VGCL** (Xia et al., 2022) for reference. (2) We denote the backbone of SIL as **Base** and combine it with other pre-training techniques, including **Wav2vec2.0** model (Baevski et al., 2020) that performs self-supervised training, **VILT** (Kim et al., 2021) that follows multi-modal pre-training, and **LUT** (Dong et al., 2021). For these pre-training techniques, we adopt the same 960h LibriSpeech data. (3) Besides, we combine our semantic task $\mathcal{L}_{\text{sem}}$ with the above text-based grounding backbones for a better comparison. We keep almost all architecture as same as the original backbones but replace the text encoder with the speech encoder.

Main Results. Table 1 reports the performance evaluation results. We also fine-tune the speech encoder on the grounding data and report the results

Table 1: Performance Evaluation on the ActivityNet Speech. The best results are bold. SP: Speech. TXT: Text.

Setting	Backbone	Pre-Training & Data			R@1,IoU=m				R@5,IoU=m				mIoU(finetune)
		Method	SP	TXT	0.1	0.3	0.5	0.7	0.1	0.3	0.5	0.7	
Supervised Approach													
E2E	VSLNet	-	×	×	76.42	49.64	31.98	17.26	-	-	-	-	35.92
	VGCL	-	×	×	75.26	51.80	32.36	18.10	-	-	-	-	36.83
Weakly-supervised Approach													
Cascaded	ASR-WSLLN	-	×	×	74.47	41.76	22.62	11.03	-	-	-	-	31.42(31.68)
	ASR-SCN	-	×	×	71.35	46.64	28.09	13.26	89.55	71.32	55.74	31.14	32.02(32.49)
	ASR-RTBPN	-	×	×	74.62	47.75	28.80	13.44	92.31	77.52	61.88	32.37	32.88(33.26)
E2E	Base	-	×	×	65.62	41.99	24.20	11.02	87.28	72.37	55.83	28.08	28.44
	Base	Wav2vec2.0	✓	×	65.65	44.61	26.20	10.83	90.01	74.21	58.37	28.89	30.22(30.36)
	Base	VILT	✓	✓	69.71	45.69	26.68	12.43	92.51	76.80	59.56	30.91	31.06(32.13)
	Base	LUT	✓	✓	72.44	46.24	26.97	12.82	92.08	75.64	60.38	31.12	31.34(32.48)
	WSLLN	$\mathcal{L}_{\text{sem}}$	✓	✓	**75.11**	41.39	22.07	10.96	-	-	-	-	31.28(31.54)
	RTBPN	$\mathcal{L}_{\text{sem}}$	✓	✓	72.86	46.76	27.89	13.24	92.49	76.80	61.04	32.45	32.10(33.01)
E2E(Ours)	Base	$\mathcal{L}_{\text{sem}}$	✓	✓	73.11	46.39	27.07	13.06	92.10	76.12	60.24	31.24	31.88
	Base	$\mathcal{L}_{\text{ASP}}$	✓	✓	74.88	48.14	28.68	13.95	93.44	79.32	61.52	32.17	33.04
	Base+$\mathcal{L}_{\text{AVCL}}$	$\mathcal{L}_{\text{ASP}}$	✓	✓	71.79	**49.46**	**30.26**	**15.22**	**94.87**	**82.28**	**63.73**	**35.48**	**34.02(34.52)**

Table 2: The Comparison of Average Inference Latency and Grounding Performance. The batch size is set to 1.

Method	mIoU	Latency	Speedup
ASR-Base	32.13	0.090s	1.00x
ASR-RTBPN	32.88	0.094s	0.95x
SIL(Ours)	**34.02**	**0.044s**	**2.04x**

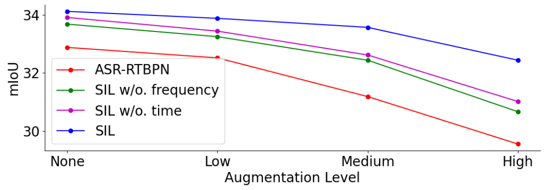


Figure 4: Robustness Analysis.

in the last column. From the results, we can observe several interesting points: (1) The cascaded baselines generally perform better than E2E baselines. This phenomenon is reasonable since the ASR model has a high recognition accuracy (see Appendix) and weakly-supervised text-based video grounding methods are mature, validating the difficulty of direct modeling the video-speech interaction. (2) In E2E baselines, our semantic task ($\mathcal{L}_{\text{sem}}$) outperforms other pre-training approaches, indicating its effectiveness for semantic modeling. However, existing text-based backbones are still inferior to their cascaded version, indicating that speech encoding and video-speech interaction are still insufficient. (3) Our full framework SIL (Base+$\mathcal{L}_{\text{AVCL}}$+$\mathcal{L}_{\text{ASP}}$) surpasses all weakly-supervised baselines on almost all criteria, verifying its effectiveness. (4) We also observe an interesting point that the stronger weakly-supervised model tends to obtain a higher value at R@1, IoU=0.7/0.5 while a lower value at R@1, IoU=0.1. This is because inaccurate models tend to give a proposal prediction with a larger temporal range, which usually covers the ground truth segment but with a low IoU score.

Grounding Efficiency. We measure the grounding efficiency in the average time required to ground one spoken query in the video. Note the computation cost of data pre-processing is excluded. We select ASR-Base/ASR-RTBPN as two typical cascaded methods for comparison and list the result in Table 2. It demonstrates the superiority of SIL in both performance and efficiency.

Robustness analysis. To showcase the robustness of our model in real-life scenes, we add augmentation at different levels (low, medium, high), including time stretching, pitch shifting and noise addition. The full configuration is introduced in Appendix B.2. We compare our SIL with ASR-RTBPN and two ablation methods that only uses frequency- or time-based features without training of robustness task. As illustrated in Figure 4, our SIL shows stable performance, validating the power of our robustness task.

4.5 Ablation Study

Overall Ablation Study. We conduct the overall ablation study and show the result in Table 3. We observe our semantic task ($\mathcal{L}_{\text{sem}}$) im-

Table 3: Ablation Results on the ActivityNet Speech.

ASP Module			AVCL Module	R@1, IoU=m				R@5, IoU=m				mIoU
$\mathcal{L}_{\text{rob}}$	$\mathcal{L}_{\text{conc}}$	$\mathcal{L}_{\text{sem}}$	$\mathcal{L}_{\text{AVCL}}$	0.1	0.3	0.5	0.7	0.1	0.3	0.5	0.7	
				67.94	41.48	22.99	10.07	88.10	72.12	56.24	29.24	28.52
		✓		73.11	46.39	27.07	13.06	92.10	76.12	60.24	31.24	31.88
	✓	✓		**74.36**	47.64	28.39	13.81	93.12	78.96	61.13	32.03	32.93
		✓	✓	72.69	47.61	29.12	14.40	92.76	79.07	62.36	33.64	32.96
	✓	✓	✓	71.13	49.37	30.02	15.19	94.56	81.32	63.38	35.27	33.91
✓	✓	✓	✓	71.79	**49.46**	**30.26**	**15.22**	**94.87**	**82.28**	**63.73**	**35.48**	**34.02**

Table 4: Evaluation Results of Different Amount of Data for Acoustic-semantic Pre-training.

Data	R@1, IoU=m			mIoU
	0.3	0.5	0.7	
10%	46.71	28.22	13.33	31.71
50%	49.28	29.42	14.46	33.87
100%	49.46	30.26	15.22	34.02

Table 5: Effect of Conciseness Task.

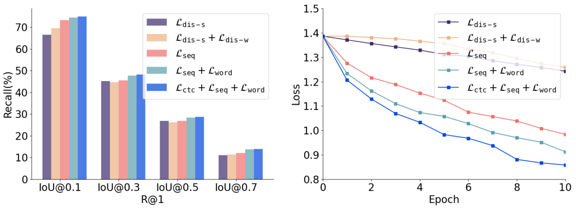
Fusion	Pre-Training	R@1, IoU=m		mIoU
		0.5	0.7	
Pool+Add	w/o. $\mathcal{L}_{\text{conc}}$	27.49	13.42	32.36
Pool+Add	w. $\mathcal{L}_{\text{conc}}$	27.24	13.38	32.28
Attention	w/o. $\mathcal{L}_{\text{conc}}$	27.07	13.06	31.88
Attention	w. $\mathcal{L}_{\text{conc}}$	**28.68**	**13.95**	**33.04**

proves performance significantly, e.g. 4.08 on R@1,IoU=0.5 and 2.99 on R@1,IoU=0.7, since it directly learns knowledge from word embedding. Also, the conciseness task ($\mathcal{L}_{\text{conc}}$) and acoustic-visual contrastive learning ($\mathcal{L}_{\text{AVCL}}$) further bring gains. Though the robustness task ($\mathcal{L}_{\text{conc}}$) has no significant effect, we have demonstrated its contribution in the robustness analysis.

4.5.1 Analysis of ASP

Effectiveness under Low-resource Setting. As shown in Table 4, under a low-resource setting, our acoustic-semantic pre-training still achieves comparable performance with limited data (50%), showing excellent generalization capability.

Effect of Conciseness Task. To investigate whether the conciseness task can promote cross-modal interaction, we combine it with two different fusion methods: attention-based fusion (Zhang et al., 2019) and simple fusion that includes pooling and addition. We remove the AVCL module to better reflect the impact and list the results in Table 5. It is observed that directly applying attention leads to inferior results due to insufficient interaction. However, our conciseness task successfully leverages the potential of the attention mechanism, improving interaction significantly for grounding.



(a) Performance (b) Grouding Loss

Figure 5: Effect of Semantic Task.

Table 6: Extension of AVCL on Text-based Grounding.

Method	R@1, IoU=m		R@5, IoU=m	
	0.3	0.5	0.3	0.5
RTBPN	49.77	29.63	79.89	60.56
RTBPN+AVCL	**50.87**	**30.96**	**84.27**	**64.68**

Effect of Semantic Task. We compare three objectives in our semantic learning task with the method (Dong et al., 2021), which also includes two similar semantic objectives: the seq-level loss $\mathcal{L}_{\text{dis-s}}$ and the word-level loss $\mathcal{L}_{\text{dis-w}}$. The difference is that they mainly focus on the distance loss (e.g. MSE) to minimize the semantic gap. The result in Figure 5 (a) reveals that our sequence-level objective $\mathcal{L}_{\text{seq}}$ can outperform $\mathcal{L}_{\text{dis-s}} + \mathcal{L}_{\text{dis-w}}$ due to the effectiveness of contrastive learning. And our word-level objective $\mathcal{L}_{\text{word}}$ can also be seamlessly combined to improve performance by a large margin. Besides, we draw the curve of grounding loss $\mathcal{L}_{\text{base}}$ during grounding training in Figure 5 (b) to reflect the video-speech alignment. We find each loss in semantic task speeds up convergence for weakly-supervised grounding training.

4.5.2 Analysis of AVCL

Integration with Text-based Video Grounding. AVCL is also a common module that can be integrated into the weakly-supervised text-based video

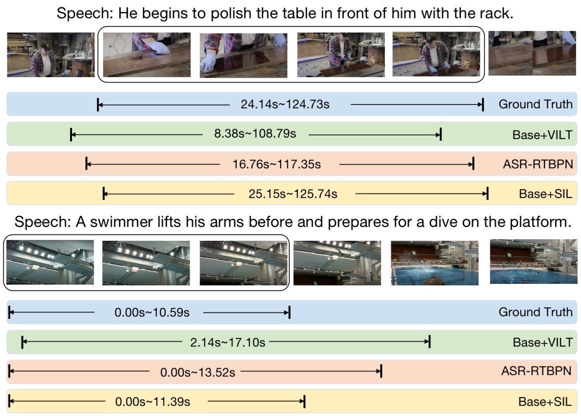
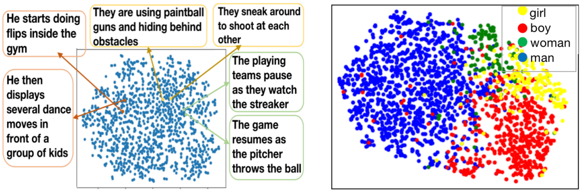


Figure 6: Visualization of Grounding Results.



(a) Overall Visualization

(b) Cluster

Figure 7: Visualization of Speech Representations.

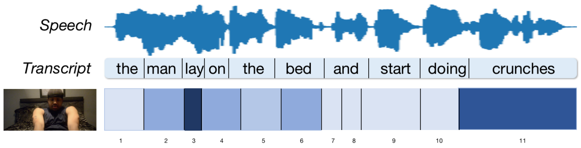


Figure 8: Visualization of Video-Speech Attention.

grounding by replacing the speech features with the word features. As shown in Table 6, our AVCL module further improves the performance on the strong basis of the RTBPN method, validating its versatility and effectiveness.

4.6 Qualitative Analysis

Visualization of Grounding Results. Figure 6 depicts two grounding examples from the ActivityNet Speech, where Base+SIL localizes the temporal moments covering the most salient part. Compared to two methods Base+VILT and ASR-RTBPN, our SIL can localize more precise moments and achieve better performance, validating its effectiveness.

Visualization of Speech Representations. To qualitatively verify our acoustic-semantic pre-training strategy, we use the pre-trained encoder to extract the features of speech in the ActivityNet Speech and visualize them using t-SNE in Figure 7 (a). We show three point pairs staying close in the feature space. The two red points on the left both describe the "jump" action, and the two yellow points on the top have similar "gun" and "hide" meanings. Note each pair contains a few similar words, indicating the close distance is determined by semantic rather than acoustic information. Also, we perform clustering with respect to four specific words in Figure 7 (b). We observe there is a clear boundary and symmetric relationship between the four clusters. The above result demonstrates the effectiveness of our pre-training strategy.

Visualization of Video-Speech Attention. We visualize the video-speech attention between the target frame and segments in Figure 8 using a thermodynamic diagram, where the darker color means a higher correlation and the temporal correspondence between the transcript and speech is also shown. From the result, we observe that the frame can attend to the segments temporally corresponding to keywords, e.g. "lay", "crunches", and ignore other irrelevant ones, e.g. "the", "on". This fact suggests that our conciseness task can detect the word-level segments and boost the cross-modal interaction.

5 Conclusion

In this paper, we propose a new task named Weakly-Supervised Spoken Video Grounding and present a novel framework SIL. Concretely, we conduct an acoustic-semantic pre-training to achieve effective and robust semantic encoding. Besides, we develop an acoustic-visual contrastive learning to optimize representations for cross-modal interaction. The extensive experiments demonstrate the superiority of our proposed method.

6 Limitations

In this section, we make a clear discussion of the limitation of our work. Our work mainly leverages a pre-training scheme to enhance the encoding of speech for video grounding. However, the adopted audio data (i.e. Libri Speech) for pre-training are different from the one in the grounding dataset (i.e. ActivityNet Speech). This could lead to performance degradation due to the domain gap. The findings could inspire the researchers to explore a better pre-training strategy to learn domain-invariant and effective speech representations for grounding.

7 Ethics Statement

We adopt the widely-used datasets that were produced by previous researchers. We follow all relevant legal and ethical guidelines for their acquisition and use. Besides, we recognize the potential influence of our technique, such as its application in human-computer interaction and vision-language grounding systems. We are committed to conducting our research ethically and ensuring that our research is beneficial. We hope our work can inspire more investigations for spoken video grounding and wish our framework can serve as a solid baseline for further research.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant No.2022ZD0162000, National Natural Science Foundation of China under Grant No.62222211, Grant No.61836002 and Grant No.62072397, and Yiwise.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970.

Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. 2018. Temporally grounding natural sentence in video. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 162–171.

Jingyuan Chen, Lin Ma, Xinpeng Chen, Zequn Jie, and Jiebo Luo. 2019a. Localizing natural language in videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8175–8182.

Long Chen, Chujie Lu, Siliang Tang, Jun Xiao, Dong Zhang, Chilie Tan, and Xiaolin Li. 2020. Rethinking the bottom-up framework for query-based video localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10551–10558.

Zhenfang Chen, Lin Ma, Wenhan Luo, and Kwan-Yee K Wong. 2019b. Weakly-supervised spatio-temporally grounding natural sentence in video. *arXiv preprint arXiv:1906.02549*.

Xize Cheng, Linjun Li, Tao Jin, Rongjie Huang, Wang Lin, Zehan Wang, Huangdai Liu, Ye Wang, Aoxiong Yin, and Zhou Zhao. 2023. Mixspeech: Cross-modality self-learning with audio-visual stream mixup for visual speech translation and recognition. *arXiv preprint arXiv:2303.05309*.

Grzegorz Chrupała. 2022. Visually grounded models of spoken language: A survey of datasets, architectures and evaluation techniques. *Journal of Artificial Intelligence Research*, 73:673–707.

Grzegorz Chrupała, Lieke Gelderloos, and Afra Alishahi. 2017. Representations of language in a model of visually grounded speech signal. *arXiv preprint arXiv:1702.01991*.

Leon Cohen. 1995. *Time-frequency analysis*, volume 778. Prentice hall New Jersey.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Linhao Dong and Bo Xu. 2020. Cif: Continuous integrate-and-fire for end-to-end speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6079–6083. IEEE.

Qianqian Dong, Rong Ye, Mingxuan Wang, Hao Zhou, Shuang Xu, Bo Xu, and Lei Li. 2021. Listen, understand and translate: Triple supervision decouples end-to-end speech-to-text translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12749–12759.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7).

Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee-Keong Kwoh, Xiaoli Li, and Cuntai Guan. 2022. Self-supervised contrastive representation learning for semi-supervised time-series classification. *arXiv preprint arXiv:2208.06616*.

Patrick Flandrin. 1998. *Time-frequency/time-scale analysis*. Academic press.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275.

Mingfei Gao, Larry S Davis, Richard Socher, and Caiming Xiong. 2019. Wslln: Weakly supervised natural language localization networks. *arXiv preprint arXiv:1909.00239*.

Lieke Gelderloos and Grzegorz Chrupała. 2016. From phonemes to images: levels of representation in a recurrent neural model of visually-grounded language learning. *arXiv preprint arXiv:1610.03342*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings.

David Harwath and James Glass. 2015. Deep multimodal semantic embeddings for speech and images. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 237–244. IEEE.

David Harwath and James Glass. 2019. Towards visually grounded sub-word speech unit discovery. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3017–3021. IEEE.

David Harwath, Wei-Ning Hsu, and James Glass. 2019. Learning hierarchical discrete linguistic units from visually-grounded speech. *arXiv preprint arXiv:1911.09602*.

Anne Lisa Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812.

Bertrand Higy, Lieke Gelderloos, Afra Alishahi, and Grzegorz Chrupała. 2021. Discrete representations in neural models of spoken language. *arXiv preprint arXiv:2105.05582*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2018. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*.

Jiabo Huang, Yang Liu, Shaogang Gong, and Hailin Jin. 2021. Cross-sentence temporal and semantic relations in video activity localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7199–7208.

Tao Jin, Siyu Huang, Yingming Li, and Zhongfei Zhang. 2020. Dual low-rank multimodal fusion. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 377–387.

Tao Jin and Zhou Zhao. 2021. Generalizable multilinear attention network. *Advances in Neural Information Processing Systems*, 34:9049–9060.

Jodi Kearns. 2014. Librivox: Free public domain audiobooks. *Reference Reviews*.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu. 2020. Weakly-supervised video moment retrieval via semantic completion network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11539–11546.

Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. 2018. Cross-modal moment localization in videos. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 843–851.

Guoshun Nan, Rui Qiao, Yao Xiao, Jun Liu, Sicong Leng, Hao Zhang, and Wei Lu. 2021. Interventional video grounding with dual contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2765–2775.

Andreea-Maria Oncescu, Joao F Henriques, Yang Liu, Andrew Zisserman, and Samuel Albanie. 2020. Queryd: A video dataset with high-quality textual and audio narrations. *arXiv e-prints*, pages arXiv–2011.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Karol J Piczak. 2015. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018.

Sebastiaan Scholten, Danny Merkx, and Odette Scharenborg. 2021. Learning to recognise words using visually grounded speech. In *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE.

Yijun Song, Jingwen Wang, Lin Ma, Zhou Yu, and Jun Yu. 2020. Weakly-supervised multi-level attentional reconstruction network for grounding textual queries in videos. *arXiv preprint arXiv:2003.07048*.

Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. 2019. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jie Wu, Guanbin Li, Si Liu, and Liang Lin. 2020. Tree-structured policy based progressive reinforcement learning for temporally language grounding in video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12386–12393.

Yan Xia, Zhou Zhao, Shangwei Ye, Yang Zhao, Haoyuan Li, and yi Ren. 2022. Video-guided curriculum learning for spoken video grounding. In *Proceedings of the 30th ACM International Conference on Multimedia*.

Jiahao Xun, Shengyu Zhang, Zhou Zhao, Jieming Zhu, Qi Zhang, Jingjie Li, Xiuqiang He, Xiaofei He, Tat-Seng Chua, and Fei Wu. 2021. Why do we click: visual impression-aware news recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3881–3890.

Aoxiong Yin, Zhou Zhao, Weike Jin, Meng Zhang, Xingshan Zeng, and Xiaofei He. 2022. Mlslt: Towards multilingual sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5109–5119.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*.

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. 2020a. Span-based localizing network for natural language video localization. *arXiv preprint arXiv:2004.13931*.

Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. 2020b. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12870–12877.

Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. 2022. Self-supervised contrastive pre-training for time series via time-frequency consistency. *arXiv preprint arXiv:2206.08496*.

Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. 2019. Cross-modal interaction networks for query-based moment retrieval in videos. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 655–664.

Zhu Zhang, Zhijie Lin, Zhou Zhao, Jieming Zhu, and Xiuqiang He. 2020c. Regularized two-branch proposal networks for weakly-supervised moment retrieval in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4098–4106.

Zhu Zhang, Zhou Zhao, Zhijie Lin, Xiuqiang He, et al. 2020d. Counterfactual contrastive learning for weakly-supervised vision-language grounding. *Advances in Neural Information Processing Systems*, 33:18123–18134.

Yang Zhao, Zhou Zhao, Zhu Zhang, and Zhijie Lin. 2021. Cascaded prediction network via segment tree for temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4197–4206.

Minghang Zheng, Yanjie Huang, Qingchao Chen, and Yang Liu. 2022. Weakly supervised video moment localization with contrastive negative sample mining.

This appendix contains four sections. (1) Appendix A introduces the detailed design of the base grounding network. (2) Appendix B presents the experiments details. (3) Appendix C provides additional analysis. (4) Appendix D describes the detail and insight of our technique.

A Base Grounding Network

A.1 Feature Encoder

Video Encoder. For each video V, we first extract its features by a pre-trained 3D ConvNet (Tran et al., 2015). Then we apply a linear layer to project them to the hidden dimension d and utilize the QA encoder blocks (Yu et al., 2018) to generate contextualized video representations $\bar{\mathbf{V}} = \{\bar{\mathbf{v}}_i\}_{i=1}^{n_v} \in \mathbb{R}^{n_v \times d}$, where n_v is the clip number and d is the hidden dimension.

Speech Encoder. The details of the speech encoder have been introduced in Section 3.3 in the main paper. Note we do not introduce the conciseness task in our Base model, thus the I&F algorithm is not utilized here.

A.2 Interaction

To model the video-speech interaction, we conduct an attentive aggregation (Chen et al., 2019a; Zhang et al., 2019), given by:

$$\begin{aligned} \delta_{ij} &= \mathbf{u}_m^\top \tanh(\mathbf{U}_m^1 \bar{\mathbf{v}}_i + \mathbf{U}_m^2 \bar{\mathbf{s}}_j + \mathbf{b}^m), \\ \tilde{\delta}_{ij} &= \frac{\exp(\delta_{ij})}{\sum_{k=1}^{n_c} \exp(\delta_{ik})}, \mathbf{q}_i = \sum_{j=1}^{n_c} \tilde{\delta}_{ij} \bar{\mathbf{s}}_j \end{aligned} \tag{9}$$

where $\mathbf{U}_m^1, \mathbf{U}_m^2$ are projection matrices in the clip-to-speech attention, $\mathbf{b}^m$ is the bias and $\mathbf{u}_m^\top$ is the row vector. The $\mathbf{q}_i$ is the summarized speech feature relevant to the i-th clip. Then we concatenate them and apply a Bi-GRU network to yield the multi-modal clip representations $\{\mathbf{m_i}\}_{i=1}^{n_v} \in \mathbb{R}^{n_v \times d}$.

A.3 Proposal Scorer

We follow 2D-TAN (Zhang et al., 2020b,c) to build a 2D feature map $\mathbf{F} \in \mathbb{R}^{n_v \times n_v \times d}$ as proposal features. Then we apply a 2D convolution layer to obtain the updated proposal features $\{\mathbf{h}_i\}_{i=1}^{n_p}$, where n_p is the number of the proposals. Finally, we apply a fully connected layer with sigmoid function to generate proposal scores $K = \{k_i\}_{i=1}^{n_p}$.

A.4 MIL Training

Under the weakly-supervised setting, we follow the MIL scheme to train the model. For each matched video-speech pair (V, S), we randomly select another video V' and speech S' from the training set to construct two negative pairs (V', S) and (V, S'). We compute the alignment score $f(K)$ for (V, S), and compute $f(K_{V'})$ and $f(K_{S'})$ for (V', S) and (V, S') similarly, where $f(\cdot)$ is the average of top-R proposal scores and R it set to 20. We adopt the binary cross-entropy (BCE) loss to learn the cross-modal alignment by:

$$\begin{aligned} \mathcal{L}_{base} = &- \log f(K) - \log(1 - f(K_{\bar{V}})) \\ &- \log f(K) - \log(1 - f(K_{\bar{S}}))) \end{aligned} \tag{10}$$

To stabilize the weakly-supervised training, we also add a widely-used diversity loss (Chen et al., 2019b; Zhang et al., 2020d) $\mathcal{L}_{div}$ for score distribution as:

$$\bar{k}_i = \frac{\exp(k_i)}{\sum_{i=1}^{\mathrm{R}} \exp(k_i)}, \quad \mathcal{L}_{div} = -\bar{k}_i \log(\bar{k}_i) \tag{11}$$

B Experiment Details

B.1 Dataset Details

ActivityNet Speech. The dataset (Xia et al., 2022) constructs the speech annotations by employing 58 speakers to read the original text descriptions in ActivityNet Captions (Caba Heilbron et al., 2015), which contain 28 male speakers and 30 female speakers. To guarantee the recording quality, all the speakers are required to read smoothly without a stammer. The average of each speech recording is 6.22 seconds and about 124.3 hours in total. Following the standard split in (Zhang et al., 2020b,c), there are 37,417, 17,505 and 17,031 moment-speech pairs used for training, validation and testing, respectively.

LibriSpeech. To evaluate different scenarios with respect to the amount of available training data, we use the standard LibriSpeech (Panayotov et al., 2015) division that includes 100, 460 and 960 hours training data. We report the performance of the model with 960 hours data for pre-training.

B.2 Implementation Details

Data Preprocessing. For video features, we split them into 64 clips following previous work (Zhang et al., 2020b). For speech input, we downsample the speech sequence to 1/4 of its original length. For log-mel spectrograms, we adopt a 16 kHz sampling rate, a 25 ms Hamming window, a 20 ms window stride, and 80 mel filter bands.

Table 7: The Augmentation Configuration on Wave Data of Speech. SNR is the signal-noise ratio.

Augmentation	low	medium	high
time stretching (ratio)	[0.8,1.25]	[0.6,1.67]	[0.5,2.0]
pitch shifting (step)	[-2,2]	[-4,4]	[-6,6]
noise addition (SNR)	[20,30]	[10.20]	[0,5]

Table 8: The Accuracy of ASR model on ActivityNet Speech.

Method	WER
wav2vec 2.0(adopted)	5.2817
Google API	9.5057

Model Setting. The dimension d of hidden layers is set to 256. The number N_a and N_s of Transformer encoder layers are both set to 4. In conciseness task, we follow (Dong and Xu, 2020) to adopt the scaling and tail-handling strategy. The non-auto-regressive decoder is an ordinary 4-layer Transformer decoder. In semantic task, the CTC decoder consists of a two-layer MLP, the number B of negative samples for sequence-level objective is set to 512, the number of Transformer decoder layers for word-level objective is set to 4 and the mask ratio x% is set to 50%. In acoustic-visual contrastive learning, the number T of positive/negative samples is set to 12. For training, the loss coefficients λ_1, λ_2 and λ_3 are empirically set to 1.0, 1.0 and 0.1 respectively. We adopt an Adam optimizer (Duchi et al., 2011) with the warmup updates of 7000 and 200 and the learning rate of 0.0002 and 0.0003 for pre-training and grounding training, respectively. During inference, we apply the non-maximum suppression (NMS) with a threshold 0.45 when selecting multiple proposals.

Training Step. For pre-training on the LibriSpeech, we train the encoder with loss $\mathcal{L}_{\text{ASP}}$ for 40 epochs. For weakly-supervised spoken video grounding training on the ActivityNet Speech, we train the model with loss $\mathcal{L}_{\text{G}}$ for 10 epochs.

Experiment Configuration. The SIL is implemented using PyTorch 1.9.0 with CUDA 10.0 and cudnn 7.6.5. All the experiments are conducted on a workstation with two NVIDIA GeForce RTX 2080Ti GPU.

Data Augmentation. In the robustness analysis, we apply speech augmentation at three levels (low, medium, high), as shown in Table 7. Time stretching means the change of speed/duration of speech wave data without changing pitch. Pitch shifting means the step change of speech pitch without changing speed and duration. Noise addition means adding the noise from ESC-50 (Piczak, 2015) to original speech at the same sampling rate.

B.3 Baseline Setting

Cascaded Methods. For the ASR model, we adopt the pre-trained wav2vec2.0 model (Baevski et al., 2020). To verify the accuracy of the ASR model on ActivityNet Speech, we make a comparison with the open API of Google ASR[2] and the result is shown in Table 8, which indicates that we do select an effective ASR model. We then detail the architectures of the above weakly-supervised temporal video grounding method. **WSLLN** (Gao et al., 2019) fuses the visual proposals with the text and conducts the proposal detection and alignment simultaneously, then generates final matching scores. **RTBPN** (Zhang et al., 2020c) builds an enhanced visual stream and a suppressed visual stream based on a language-guided filter, then fuses them with text and considers both the intra-sample and inter-sample loss to train the model with additional regularization terms. It also adopts 2D-TAN (Zhang et al., 2020b) as the backbone for proposal generation and modeling. **SCN** (Lin et al., 2020) utilizes a Transformer decoder (Vaswani et al., 2017) to reconstruct the masked language based on the visual proposals and rank them based on the reward.

- **Performance of Original Text-based Methods.** To clearly show the effect of ASR on the grounding, we also present the performance of original text-based methods for reference. As shown in Table 9, the performance of cascaded methods is inferior to the original methods, especially on some strict criteria on R@1, e.g. ASR-RTBPN drops 2.02 on R@1,IoU=0.3 and 0.83 on R@1, IoU=0.5. Meanwhile, we also observe stronger baselines suffer from more severe degradation, suggestting the limitations of cascaded methods even with an excellent ASR model. Further, our proposed framework SIL is able to achieve comparable even better results compared with original text-based methods, e.g. 49.46 vs 49.77 on R@1,IoU=0.3 and 30.26 vs 29.63 on R@1,Iou=0.5.

End-to-end Methods. We next introduce the detailed architectures of the end-to-end baselines:

- **Supervised Methods. VSLNet** (Zhang et al., 2020a) is originally proposed for text-based

[2] https://cloud.google.com/speech-to-text.

Table 9: Performance Comparison Results on the ActivityNet Speech. The values in bracket denote the performance of original text-based methods (when accuracy of ASR is 100%).

Method	R@1,IoU=m			R@5,IoU=m		
	0.1	0.3	0.5	0.1	0.3	0.5
ASR-WSLLN	$74.47_{(75.40)}$	$41.76_{(42.80)}$	$22.62_{(22.70)}$	-	-	-
ASR-SCN	$71.35_{(71.48)}$	$46.64_{(47.23)}$	$28.09_{(29.22)}$	$89.55_{(90.88)}$	$71.32_{(71.45)}$	$55.74_{(55.69)}$
ASR-RTBPN	$74.62_{(73.73)}$	$47.75_{(49.77)}$	$28.80_{(29.63)}$	$92.31_{(93.89)}$	$77.52_{(79.89)}$	$61.88_{(60.56)}$
SIL(Ours)	71.79	49.46	30.26	94.87	82.28	63.73

video grounding. It directly estimates the frame-level probabilities of being boundaries, where a cross-entropy loss is utilized to supervise the probability distribution. We simply replace the textual features with the log-mel spectrograms features of speech. **VGCL** (Xia et al., 2022) is directly proposed for spoken video grounding. It utilizes matched frame-level features to perform contrastive predictive coding for speech encoding and then conduct grounding training.

- **Weakly-supervised Methods.** The **Base** adopts 80-dimensional log-mel spectrograms as the speech input. Without pre-training, we reduce the layer number of Transformer encoder to 2 for speech encoding, which achieves better and more stable performance. The **Base+Wav2vec2.0** adopts the 512-dimensional vectors obtained from the last layer of pre-trained wav2vec2.0 (Baevski et al., 2020) model as the input speech features. It also follows the Base architecture and reduces the layer number of Transformer encoder to 2 for stability. The **Base+VILT** follows the text-video multi-modal pre-training strategy (Kim et al., 2021) to build an 8-layer cross-modal Transformer encoder encoding both text and speech modalities. It develops two tasks, where the text-speech matching identifies whether this textual sentence corresponds to the speech and the language modeling randomly masks 15% of words for prediction. We adopt the same word embedding and speech input features as our SIL. We first pre-train the encoder with 30 epochs and fix it. During grounding training, we apply another trainable fully-connected layer with a Bi-GRU to encode speech and adopt the Base architecture. The **Base+LUT** follows the speech recognition pre-training (Dong et al., 2021) to build an 8-layer Transformer encoder for speech. It develops the knowledge distillation to learn word features at sequence-level and word-level with MSE loss. We adopt the same word embedding and speech input features as our SIL. We first pre-train the encoder with 30 epochs and fix it, then similarly adopt the Base architecture for grounding. The **WSLLN+$\mathcal{L}_{\mathbf{sem}}$** and **RTBPN+$\mathcal{L}_{\mathbf{sem}}$** both utilize our semantic task to pre-train the speech encoder while adopt the model architectures of WSLLN and RTBPN for grounding training, respectively. Similar to SIL, we pre-train the speech encoder and fix it during grounding training. Note the reconstruction-based methods utilizing the word reconstruction can't be applied to speech, thus we do not combine our semantic task with SCN. The **Base+$\mathcal{L}_{\mathbf{sem}}$** is our ablation model, with only the semantic task as pre-training.

C Additional Analysis

C.1 Effect of ASP under Supervised Setting.

To better reflect the performance of our acousitc-semantic pre-training, we further apply it to the supervised approaches under the end-to-end framework. We first build two baselines adopting log-mel spectrograms as speech input features without any pre-training for the speech encoder: **VSLNet** (Zhang et al., 2020a) and **SBase**. The SBase is the same as Base in Appendix B.3, but we train it under the supervised setting. Here we follow (Zhang et al., 2020b) to calculate the IoU between each proposal and ground truth as the supervision for the proposal score and utilize a binary cross-entropy loss to train the model. Besides, we also report the performance of **VGCL** (Xia et al., 2022) which utilize matched video clips to perform contrastive predictive coding pre-training for speech. Then we apply our acousitc-semantic pre-training on the SBase model to pre-train the speech encoder and then conduct supervised grounding training. As shown in Table 10, our supervised baseline SBase is inferior to VSLNet and VGCL. However, with our acoustic-semantic pre-training strategy, the approach SBase+ASP significantly

Table 10: Effect of ASP module under Fully-Supervised Setting. FS: fully-supervised.

Method	Setting	Pre-Training	R@1,IoU=m				R@5,IoU=m				mIoU
			0.1	0.3	0.5	0.7	0.1	0.3	0.5	0.7	
VSLNet	FS	-	76.42	49.64	31.98	17.26	-	-	-	-	35.92
VGCL	FS	VGCL	75.26	51.80	32.36	18.10	-	-	-	-	36.83
SBase	FS	-	69.10	48.70	30.53	15.38	96.62	86.53	74.82	46.19	33.14
SBase	FS	ASP	79.02	59.61	40.32	21.00	95.18	86.11	76.10	49.26	40.54

Table 11: Effect of Sampling Strategies in AVCL.

Method	R@1, IoU=m			
	0.1	0.3	0.5	0.7
SIL w/o. location	72.38	49.14	29.51	14.71
SIL w/o. score	**72.62**	48.96	29.28	14.42
SIL	71.79	**49.46**	**30.26**	**15.22**

outperforms other baselines, demonstrating the effectiveness of our ASP module for semantic learning even under the supervised setting.

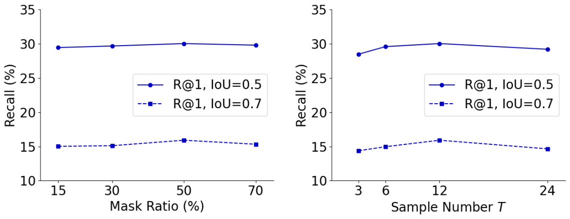
C.2 Effect of Sampling Strategies in AVCL.

To evaluate two proposed sampling strategies in our acoustic-visual contrastive learning, we generate two ablation models **SIL w/o. location** and **SIL w/o. score**. As shown in Table 11, removing each sampling strategy will lead to performance drop. We also note that score-based mining has a larger impact than location-based selection. This is because the former can more effectively filter out hard negative samples, leading to a more precise optimization for cross-modal interaction.

C.3 Hyper-parameter Analysis

Impact of Mask Ratio x%. We set the mask ratio x% to [15%, 30%, 50%, 70%] to explore its impact. We display the results in Figure 9 (a). We note that the performance first gradually improves and then slowly decreases with the increase of mask ratio. When the mask ratio is set to 50%, the model can achieve best performance. This phenomenon is slightly different from the setting reported in BERT (Devlin et al., 2018), where the ratio is set to a small value 15%. This is because a larger mask ratio in the word-level objective empowers the decoder to predict the masked words based on the speech input, reducing the dependence on contextual features.

Positive/Negative Sample Number T. Since the total clip number is set to 64, we set T to [3, 6, 12, 24] to study its effect. As shown in Figure 9 (b), the performance first improves then decreases as T increases and the AVCL module performs best when T is set to 12. This observation suggests that the limited samples fail to provide the encoder with sufficient and discriminative features necessary for effective distinction, while an excessive number of samples introduce noisy and inaccurate features that hamper performance.



(a) Mask Ratio x% (b) Sample Number T

Figure 9: Hyper-parameter Analysis of the Mask Ratio x% and the Sample Number T.

D Technique Details and Insight

D.1 Insight of Robustness Task

Speech signals, as time series data, are easily biased by complex factors, such as large variations of temporal dynamics across datasets, real-life noise and irregular sampling (Zhang et al., 2022). Thus, the speech encoder is required to be robust to extract valuable information from the speech, which ensures stable training for the downstream weakly-supervised spoken video grounding task.

To enhance the robustness of speech encoding, our goal is to identify a general property that remains consistent across diverse speech sequences. Recent findings (Zhang et al., 2022) highlight the advantages of a latent time-frequency space for representation learning. By decomposing the speech signal into the frequency and time domains, we can view them as complementary perspectives of the same data (Cohen, 1995). This relationship, rooted in signal processing theory, provides an inherent invariance that persists regardless of the distribution of time series (Flandrin, 1998; Eldele et al., 2022), serving as an inductive bias for pre-training.

As the consistency loss used in computer vision to learn invariant features across different transformations (e.g., rotation, translation, scaling), it is both reasonable and necessary to explore the time and frequency domains in speech signals and enforce the inter-domain consistency of encoded features, ensuring the underlying invariance properties (i.e., semantics) are captured during pre-training.

D.2 Integrate and Fire Method

Integrate and Fire (Dong and Xu, 2020) is originally proposed for ASR field, we utilize it to automatically detect the boundary and extract segment-level features for the conciseness task. By inserting it before the semantic learning, we can ensure each segment serve as an independent semantic unit such as a word and hence reduce the redundant information of original long sequence that may cause over-smooth cross-modal attention distribution.

As mentioned in Section 3.3.2, first, the input acoustic sequence $\mathbf{S}_a = \{\mathbf{s}_{a,i}\}_{i=1}^{n_s}$ will be fed to a weight predictor consisting of a multi-layer perceptron to obtain the weights $G = \{g_i\}_{i=1}^{n_s}$, representing the amount of information in $\mathbf{S}_a$. Then the I&F method scans and accumulates them from left to right until the sum reaches the threshold θ (set to 1.0), indicating a semantic boundary b_j is detected. Third, the current scanned weight g_{b_j} will be split into two parts: l_j and r_j. The l_j is used for fulfilling the integration of the current segment $s_{c,j}$ while r_j is used for the next integration of $s_{c,j+1}$. Then, the I&F method resets the accumulation and continues to scan the rest which begins with r_j. Finally, we multiply all g_i by corresponding $\mathbf{s}_i$ and integrate them based on detected boundaries to obtain segment-level features $\mathbf{S}_c = \{\mathbf{s}_{c,i}\}_{i=1}^{n_c}$, where n_c is the number of segments. The process can be formulated as follows:

$$
\begin{aligned}
&b_j = \underset{t}{argmin}(r_{j-1} + \sum\nolimits_{i=b_{j-1}+1}^{t} g_i > \theta), \\
&r_j = r_{j-1} + \sum\nolimits_{i=b_{j-1}+1}^{b_j} g_i - T, l_j = g_{b_j} - r_j, \\
&\mathbf{s}_{c,j} = r_{j-1} * \mathbf{s}_{a,b_{j-1}} \\
&\qquad + \sum\nolimits_{i=b_{j-1}+1}^{b_j} g_i * \mathbf{s}_{a,i} + l_j * \mathbf{s}_{a,b_j}
\end{aligned}
\tag{12}
$$

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 6.

☒ A2. Did you discuss any potential risks of your work?
It has no obvious risks.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
No response.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

Section 4.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix B.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4 and Appendix B.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix B.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.