Localizing Moments in Long Video Via Multimodal Guidance

Wayner Barrios[1], Mattia Soldan[2], Alberto Mario Ceballos-Arroyo[3],
Fabian Caba Heilbron[4], Bernard Ghanem[2]

[1] Dartmouth [2]King Abdullah University of Science and Technology (KAUST) [3]Northeastern University [4]Adobe Research

Abstract

The recent introduction of the large-scale, long-form MAD and Ego4D datasets has enabled researchers to investigate the performance of current state-of-the-art methods for video grounding in the long-form setup, with interesting findings: current grounding methods alone fail at tackling this challenging task and setup due to their inability to process long video sequences. In this paper, we propose a method for improving the performance of natural language grounding in long videos by identifying and pruning out non-describable windows. We design a guided grounding framework consisting of a Guidance Model and a base grounding model. The Guidance Model emphasizes describable windows, while the base grounding model analyzes short temporal windows to determine which segments accurately match a given language query. We offer two designs for the Guidance Model: Query-Agnostic and Query-Dependent, which balance efficiency and accuracy. Experiments demonstrate that our proposed method outperforms state-of-the-art models by 4.1% in MAD and 4.52% in Ego4D (NLQ), respectively. Code, data and MAD's audio features necessary to reproduce our experiments are available at: https://github.com/waybarrios/guidance-based-video-grounding.

1. Introduction

The task of grounding natural language in long-form videos within large-scale datasets, such as MAD [33] and Ego4D [12], can be particularly challenging due to the possibility of encountering many video segments that do not contain any interesting or relevant moments to search for. Given the large search space, it is critical to develop approaches that can quickly scan the video and identify queryable moments via natural language.

For instance, Figure 1 illustrates various moments occurring in a long-form video, where certain moments could have greater relevance than others based on the video's storytelling. The dialogue moment, which spans more than three minutes, would only be described as *"two people sitting at a table"*, whereas the action scene contains many moments that can be described, such as *"a girl falling into a pool"*, *"a dog opening a fence"*, and *"a dog jumps to rescue a little girl"*. This demonstrates that certain slices within a video can contain numerous notable moments that users would be highly interested in searching for.

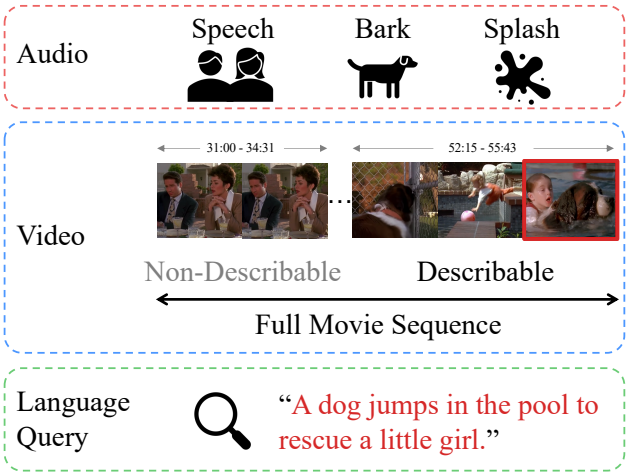

Figure 1: **Describable Window.** We depict the difference between describable and non-describable windows. The former are temporal windows with relevant visual and auditory events that are likely to contain one or more noteworthy moments. The latter can be categorized as "boring" video segments (temporal windows) where little happens and no moment of interest happens for grounding.

In consequence, we introduce the concept of *Describable windows* (as shown in Figure 1), which are video slices shorter than the original long-form video and have a high probability of containing remarkable and relevant visual moments that can be queried through natural language. Conversely, video slices that do not contain such visually-rich moments that cannot be queried through natural language are considered *non-describable windows*.

Our work builds on the hypothesis that certain video segments (*non-describable windows*) are difficult to narrate effectively by natural language, which can lead to noisy predictions when retrieving moments via natural language queries, particularly for long videos. In fact, state-of-the-art grounding models achieve remarkable performance when analyzing short videos [31, 16, 26]; however, their capabilities seem to saturate when tested on longer videos [33]. For example, VLG-Net [34] achieves only marginal gains compared to a simple yet effective zero-shot approach based on CLIP [29], and some models outright fail to achieve reasonable performance[1]. The above statements highlight the need to detect video slices corresponding to the "important" parts of a long-form video, *i.e.*, those that contain as many moments as possible. Our intuition is that when analyzing the entire long-form video, grounding models might overlook the most critical segments of the video, which represents a gap in the current research and motivates us to develop more sophisticated techniques that can capture certain intricacies using language cues or audio cues on top of the visual information.

To address this issue, we propose a guided grounding framework [1, 10] comprised of two core components: a *Guidance Model* that specializes emphasizing describable windows and a base grounding model that analyzes short temporal windows to determine which temporal segments match a given query accurately.

Observing multimodal cues is key for detecting describable windows. For instance, let us suppose we want to find the moment when *"the dog jumps to rescue the little girl"* (Figure 1). Here, the water splash (sound) and dog jumping (visual) are hints suggesting that lots of visual activities are happening in the scene. This intuition motivates the multimodal design of our Guidance Model. In practice, our model jointly encodes video and audio over an extended temporal window using a transformer encoder [35].

The proposed Guidance Model can also be used in two different frameworks: *Query Agnostic* and *Query Dependent*. The former pre-computes which parts of a video have a low-probability of containing describable windows, making it suitable for real-time applications or limited computational resources. In contrast, the latter provides more precise results by identifying irrelevant parts of a video based on a given text query, at the cost of being more computationally expensive as the number of queries grows. We also highlight the fact that the two-stage approach can be adapted to any grounding method, making it a flexible and versatile option for a wide range of applications.

[1]We trained Moment-DETR [17] from scratch on the MAD [33] dataset and found that, even though it achieves good grounding performance in short videos, it utterly fails in long videos. More details can be found in the supplementary material.

In summary, our **contributions** are three-fold:

1. We propose a two-stage guided grounding framework: a general approach for boosting the performance of current state-of-the-art grounding models in the long-form video setting.

2. We introduce a Guidance Model that is capable of detecting describable windows: a temporal window that contains one or more noteworthy moments.

3. Through extensive experiments, we empirically show the effectiveness of our two-stage approach. In particular, we validate the benefits of leveraging a Guidance Model that specializes in finding describable windows. We improve state-of-the-art performance on the MAD dataset [33] by 4.1% and improve the performance of an Ego4D baseline model [12, 44] in the NLQ task by 4.52%.

2. Related Work

Video Grounding Methods. Video grounding methods can be divided into two families of approaches: **(i)** proposal-based [11, 1, 34, 45, 40] and **(ii)** proposal-free [17, 43, 26, 6, 32, 19]. Proposal-based methods rely on producing confidence scores or alignment scores for a previously generated set of M candidate temporal moments $\{(\tau_{start}, \tau_{end})\}_1^M$. On the other hand, proposal-free methods aim at directly regressing the temporal interval boundaries for a given video-query pair. For instance, Mun *et al*. [26] tackle this problem by using a temporal attention-based mechanism that exploits local and global information in bimodal interactions between video segments and semantic phrases in the query to regress the target interval. Li *et al*. [19] uses a pyramid network architecture to leverage multi-scale temporal correlation maps of query-enhanced video features as the input to a temporal-attentive regression module.

Regarding efficiency, proposal-free approaches have faster inference time since they do not require exhaustively matching the query to a large set of proposals. Conversely, proposal-based methods often provide higher performance at the cost of a much slower inference time. In fact, these methods produce predictions for a large number of temporal proposals and successively apply expensive Non-Maximum Suppression (NMS) algorithms to improve the ranking.

Soldan *et al*. [33] showcased how current state-of-the-art grounding methods fail to tackle the long-form grounding setup where videos are up to several hours in duration. Our focus is thus to design a pipeline able to boost these methods through a two-stage cascade approach. In the experimental section, we will showcase how our flexible design is able to boost both proposal-based and proposal-free methods to achieve a new state-of-the-art performance.

Long-form Video Grounding. In long-form video understanding, grounding approaches are limited by the inability to process the entire video at once. Such limitation stems from the large computational budget for deep learning-based video perception algorithms and the limited computational resources currently available. Recently, [33] proposed to process the video in short windows to enable grounding in long-form videos. Long videos are therefore sliced in overlapping temporal windows, and grounding methods are used to generate predictions within each window. Finally, predictions across all windows are gathered and ranked according to the confidence scores.

While video slicing is a potential solution for long-form video grounding, it has a significant limitation as grounding methods may introduce a considerable number of false positives in the predictions for any long-form videos. We address this challenge by introducing a two-stage approach that guides any grounding model in making accurate predictions and reducing the presence of false positives.

Multimodal Transformers. The great flexibility of the transformer architecture [36] has promoted its adoption in several different fields: computer vision [9, 22], natural language processing [8, 3, 38], and audio processing [37, 27]. Concurrently, several works have exploited this architecture to process multimodal data [25, 29, 18, 20, 39, 5, 15, 30]. The key to successfully applying transformers to multimodal data is to learn a projection to a shared embedding space for each element of each modality (frames, spectrograms, language tokens). Our design for the Guidance model, therefore, follows these recent advancements. In particular, we aim to exploit the flexibility of transformers to design a unified architecture able to gather relevant cues from different modalities. We detail the design choices for our Guidance model in the following section.

Temporal Proposals. Temporal proposals refer to the identification and localization of specific actions or events in an untrimmed video by identifying temporal segments that are likely to contain them. [46, 4] generate a single video snippet per each specific event followed by an action classifier, which makes it dependent on the label class. The existence of several video snippets that correspond to specific events in a long-form video, where multiple events can occur simultaneously, significantly adds to the complexity of the task. For a given time segment t, N video snippets would be required, where N represents the number of concurrent events.

Recently, [41, 13, 28] leverages mechanisms that capture semantic information of video clips, self-attention, as well as local and global temporal relationships using visual information. Although these methods have yielded outstanding outcomes, the nature of grounding in long-form videos necessitates an examination of supplementary cues from other modalities to identify as many events as possible that can be referenced in a shorter video duration, i.e., in a describable window.

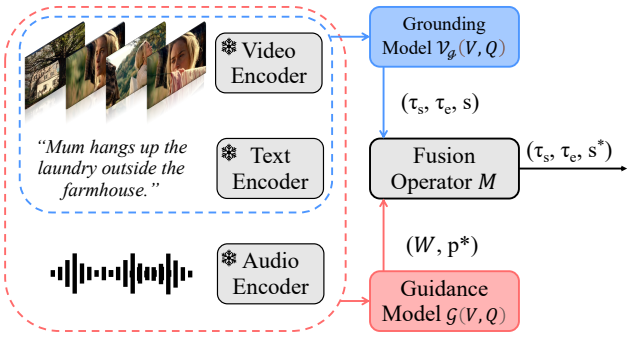

Figure 2: **Guided Grounding pipeline.** Two-stage approach comprised of a video-language grounding model and a guidance model. Given a set of predictions $\{(\tau_s, \tau_e, s)\}_1^M$ from a grounding model, and plausibility score p^*, we generate a refined set of predictions $\{(\tau_s, \tau_e, s^*)\}_1^M$.

3. Method

Our goal is to train a guidance model $\mathcal{G}$ that finds describable windows that might contain a moment to be grounded in a long video. By finding these windows, we aim to guide an existing grounding model $\mathcal{V}_g$ in improving its predictions. Let $\mathcal{V}_g$'s temporal moments predictions be defined as: $\{(\tau_s, \tau_e, s)\}_1^M$, where (τ_s, τ_e) corresponds to start/end time and s is the noisy confidence score; given a temporal window W sampled from a long video, and the fusion operator $\mathcal{M}$, we seek a model that generates temporal moment predictions such that:

$$\mathcal{M}\left(\mathcal{V}_g, \mathcal{G}, W\right) \rightarrow \{(\tau_s, \tau_e, s^*)\}_1^M, \tag{1}$$

where s^* denotes a refined confidence score for the temporal moment.

Grounding model $\mathcal{V}_g$. The grounding model takes as input a video observation V sampled from a temporal window W, and a natural language query Q; it then predicts M temporal moments such that:

$$\mathcal{V}_g\left(V, Q\right) \rightarrow \{(\tau_s, \tau_e, s)\}_1^M. \tag{2}$$

In this work, we leverage several pre-trained grounding models such as [17, 33, 34, 44]. While these works achieve remarkable performance on localizing *existing* moments within short temporal window, they suffer severely from false positive predictions on windows that do not contain any moment. As we will show in our experiments, modulating the confidence scores of a grounding model can

greatly improve the overall performance of the model.

Guidance model $\mathcal{G}$. We seek to train a guidance model that gives low scores to predictions made on empty windows (w/o moments) and gives higher scores to predictions derived from windows containing one or more moments (*describable windows*). Similar to the grounding model, $\mathcal{G}$ ingests a video observation V sampled from W, and the target query Q. For a given window, we obtain a plausibility confidence score p^* as follows:

$$\mathcal{G}\left(V, Q\right) \rightarrow p^*. \tag{3}$$

Guided grounding output $\mathcal{M}$. After obtaining a set of predictions $\{(\tau_s, \tau_e, s)\}_1^M$, and a plausibility score p^* for window W, the fusion operator M aims to generate a refined confidence score for each predicted moment (Figure 2). To do so, we simply multiply the plausibility score p^* with each confidence score $\{s\}_1^M$. This generates the final (and refined) set of predictions:

$$P = \{(\tau_s, \tau_e, s^*)\}_1^M \tag{4}$$

Grounding in long-form videos. We follow [33] and generate predictions in long-form videos by sliding a short window throughout time. Assuming we analyze K windows, our model generates $K \times M$ predictions for a natural language query Q. We sort the resulting predictions by their confidence scores s^*.

3.1. Guidance Model Design and Training Details

The Guidance Model is depicted in Figure 3. Our goal is to train the model $\mathcal{G}$ on a dataset $\mathcal{D}$ containing binary labels (of window plausibility) for a set of temporal windows and their corresponding natural language queries.

Window Representation. We sample an observation V from an input window W. Such observation includes diverse inputs (*i.e.*, video, audio, and text). The use of each modality is setup-dependent and will be thoroughly ablated in Section 4.3. In practice, we employ frozen encoders to extract embeddings for each modality. Formally, we denote video embeddings, audio embeddings, and textual embeddings as $E_v \in \mathbb{R}^{L_{vg} \times D_v}$, $E_a \in \mathbb{R}^{L_{ag} \times D_a}$, and $E_t \in \mathbb{R}^{L_{tg} \times D_t}$, respectively, where D_v, D_a, and D_t represent the dimensionality of the features. Input features are projected to a shared embedding space of size d_g by an MLP projection followed by a layer normalization module [14]. Dropout is used for regularization.

Our design allows our Guidance Model to be query-dependent or query-agnostic. The inputs for each of these variants are as follows:

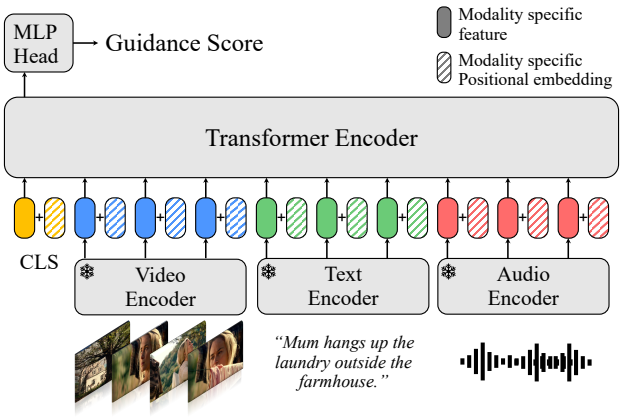

Figure 3: **Guidance model.** Our guidance model can process different modalities (setup-dependent). Each modality representation is provided with a modality-specific positional embedding before being fed to the Transformer Encoder module. The output of the CLS token is projected to a real value, used as a guidance score to condition the grounding models.

For **query-dependent**, the model's inputs are given by $E_{in}{=}|E_{cls}, E_v, E_a, E_t| \in \mathbb{R}^{(L+1)\times d_g}$, where $L{=}L_{vg}{+}L_{ag}{+}L_{tg}$ and E_{cls} is a learnable CLS token.

For **query-agnostic**, the model's inputs are given by $E_{in}{=}|E_{cls}, E_v, E_a| \in \mathbb{R}^{(L+1)\times d_g}$, where $L{=}L_{vg} + L_{ag}$ and E_{cls} is a learnable CLS token.

A modality-specific positional embedding is also added to each input tensor; specifically, we adopt a sinusoidal positional embedding [35] for the video modality, while we find that learnable positional embeddings [21] work best with the audio and text modalities.

Architecture. We adopt a transformer encoder [17, 15] which is a powerful yet flexible architecture for processing sequential data. The input E_{in} is fed to a stack of L_t transformer encoder layers. Each transformer encoder layer is identical in design to [35, 42, 17, 15], using a multi-head self-attention layer and a feed-forward network (FFN). Finally, the first element of the encoder output (E_{out}), corresponding to the CLS token position, is fed to an MLP for predicting the plausibility score for the given window.

Loss function and supervision definition. We optimize a binary cross entropy (BCE) loss to train $\mathcal{G}$. The query-agnostic models are trained on a dataset $\mathcal{D}_{agnostic}$ that contains only a set of windows and a binary label that indicates, for each window, whether or not it contains at least one moment. The query-dependent variant is trained on a dataset $\mathcal{D}_{dependent}$ that contains window and query pairs making up the set of describable (positives) and negative windows.

4. Experiments

Metrics. The video grounding metric of choice is Recall@K for IoU=θ (R@K-IoU=θ). This widely adopted metric measures the quality of the predictions' ranking and temporal overlapping (IoU) with the annotations. In this work we evaluate our models for $K\in\{1, 5, 10, 50, 100\}$ and $\theta\in\{0.1, 0.3, 0.5\}$. Additionally, we introduce Mean Recall@K (mR@K), where we average the R@K-IoU=θ performance across all IoUs thresholds. This metric allows for easier comparison and more compact tables. We also include Mean Recall$_{all}$ or average Recall as: mR$_{all} = \frac{1}{|K|}\sum$ mR@$K, \forall K$.

Datasets. MAD [33] is a large-scale dataset for the video-language grounding task that comprises more than 384K natural queries temporally grounded in 650 full-length movies for a total of over 1.2K hours of video. MAD introduces a new long-form video grounding setup that brings new challenges to the task at hand while allowing for unprecedented bias-free performance investigation. In our experimental section, unless otherwise specified, we report performance on the official test set. Ego4D [12] is a comprehensive benchmark consisting of egocentric videos from 931 camera wearers worldwide in different scenarios. The subtask we focus on is the Natural Language Query (NLQ), which uses 13 question types to locate different types of information and retrieve relevant moments from the episodic memory of camera wearers. Videos vary in length from 3.5 to 20 minutes.

Baselines. For MAD dataset, we select three video grounding methods for our experimental setup, namely: VLG-Net [34], zero-shot CLIP [33], and Moment-DETR [17]. The first two are both proposal-based methods, a trained method and a zero-shot one, while the third is a proposal-free approach. On the other hand, for Ego4D, we use Moment-DETR [17] and VSL-Net [44], a multimodal span-based framework based on context-query attention. We will show that all approaches benefit from the score refinements provided by our Guidance Model.

4.1. Implementation details

Feature Extraction. The visual and text embeddings are extracted following the CLIP-based methodology presented in [33]. Specifically, visual features are extracted at 5 FPS and with embedding dimensionality D_v=512. On the other hand, text features are comprised of N tokens, with embedding dimensionality D_t=512. Finally, audio embeddings are computed using OpenL3 [7], an audio embedding model inspired by L3-Net [2]. Specifically, we use the OpenL3 checkpoint that was pre-trained on videos containing environmental audiovisual data, we use a spectrogram time-frequency representation with 128 bands, and we set the audio embedding dimensionality D_a to 512. Furthermore, we extract the audio embeddings using a stride size equal to 0.2 seconds, *i.e.*, using an extraction frame rate of 5 Hz, which matches the frame rate of the visual features. On the other hand, in the Ego4D experiment we follow [23] without using audio streams, as not all videos contain audio throughout their entire duration.

Guidance Model. We train our Guidance Model using three modalities: **(i)** visual, **(ii)** audio, and **(iii)** text for *query-dependent* setup. And two modalities: **(i)** visual and **(ii)** audio for *query-agnostic* setup. Our transformer encoder comprises 6 layers (L_t) with a hidden size of 256 (d_g). The Guidance Model is trained using a sliding window approach where the window size (L_{vg}) is equal to 64 frames (unless otherwise specified), spanning 12.8 seconds for the MAD dataset and 34.13 seconds for the Ego4D dataset. The model is trained for 100 epochs using the AdamW [24] optimizer with learning rate 10^{-4}, weight decay 10^{-4}, and batch size 512.

Grounding Models. We instantiate Moment-DETR [17] using a hidden dimension size of 256, 2 encoder layers, 2 decoder layers, window length (L_v) of 128 (equivalent to 25.6 seconds at 5 FPS in MAD), and 10 moment queries; we optimize the model with AdamW [24], setting the learning rate to 10^{-4} and the batch size to 256. VLG-Net [34] is trained following the implementation details presented in [33], while CLIP is evaluated in the zero-shot setup that was also proposed in [33]. For VSL-Net [44], we follow the approach presented in the Ego4D [12] repository with an egocentric video-language pretraining [23]. At inference time, we discard highly redundant proposals via non-maximum suppression (NMS) with a threshold of 0.3 for MAD and 0.5 for Ego4D. All experiments are conducted on a Linux workstation with a single NVIDIA 32GB V100 GPU and an Intel Xeon CPU with 64 cores.

4.2. Results

Table 1 presents the performance of the baselines (rows 1-3) and the corresponding improvement achieved by the grounding models with the guidance model (rows 4-6). The employed guidance model utilizes query-dependent setup and audiovisual features. Further investigation into various model design choices can be found in Section 4.3.

As anticipated, proposal-based methods (rows 1-2) have consistently higher performance with respect to the proposal-free Moment-DETR method (row 3) for all metrics. Moreover, our guidance method is able to boost the performance of all baselines (rows 4-6). In particular, the most significant boost is obtained by Moment-DETR, for which the improvement ranges between $3-16\times$, while for VLG-Net and zero-shot CLIP baselines, the improve-

Model	IoU=0.1					IoU=0.3					IoU=0.5				
	R@1	R@5	R@10	R@50	R@100	R@1	R@5	R@10	R@50	R@100	R@1	R@5	R@10	R@50	R@100
Zero-shot CLIP [33]	6.57	15.05	20.26	37.92	47.73	3.13	9.85	14.13	28.71	36.98	1.39	5.44	8.38	18.80	24.99
VLG-Net [34]	3.50	11.74	18.32	38.41	49.65	2.63	9.49	15.20	33.68	43.95	1.61	6.23	10.18	25.33	34.18
Moment-DETR [17]	0.31	1.52	2.79	11.08	19.65	0.24	1.14	2.06	7.97	14.29	0.16	0.68	1.20	4.71	8.46
†Zero-shot CLIP [33]	**9.30**	**18.96**	24.30	39.79	47.35	**4.65**	13.06	17.73	32.23	39.58	2.16	7.40	11.09	23.21	29.68
†VLG-Net [34]	5.60	16.07	23.64	45.35	55.59	4.28	**13.14**	**19.86**	39.77	49.38	**2.48**	**8.78**	**13.72**	**30.22**	**39.12**
†Moment-DETR [17]	5.07	16.30	**24.79**	**50.06**	**61.79**	3.82	12.60	19.43	**40.52**	**50.35**	2.39	7.90	12.06	24.87	30.81

Table 1: **Benchmarking of grounding methods on the MAD dataset.** In the initial rows (row 1, 2, and 3), we present the performance results of the three baselines (Zero-shot CLIP, VLG-Net, and Moment-DETR) on the test split. The subsequent rows (row 4, 5, and 6) demonstrate the performance enhancement achieved by the grounding models aided by the proposed guidance model (indicated by the (†) symbol), utilizing the query-dependent setup with audiovisual features.

Method	mR@1	mR@5	mR@10	mR@50	mR@100	mR_{all}
Moment-DETR [17]	6.72	19.68	23.85	24.67	–	18.73
VSL-Net [44, 23]	11.38	19.64	24.27	36.09	42.12	26.70
†Moment-DETR [17]	7.28	22.14	24.75	26.05	–	20.05
†VSL-Net [44, 23]	**11.90**	**24.06**	**33.31**	**54.59**	**54.90**	**31.22**

Table 2: **Benchmarking on the Ego4D dataset.** The initial rows (row 1 and 2) display the performance results of Moment-DETR [17] and VSL-Net [44] baselines, employing Egocentric VLP features [23] on the Ego4D validation set. The subsequent rows (row 3 and 4) demonstrate the performance enhancement brought about by our Guidance Model, indicated by the symbol †. For a complete table of results, kindly consult the supplementary material.

ment ranges between $1-3\times$. Remarkably, the Guidance Model can boost R@10-IoU=0.1 from 2.79% to 24.79% for Moment-DETR, allowing us to achieve state-of-the-art performance for the MAD dataset. These results indicate that our approach bridges the gap between proposal-free and proposal-based methods, allowing the former to tackle the long-form video grounding task.

We evaluated the performance of the guidance model on the Ego4D [12] dataset using a query-dependent setup. Since Ego4D lacks audio in a substantial portion of its videos for the NLQ task, we opted for a fair comparison using only visual and text features. Table 2 displays the results achieved, employing VSL-Net and Moment-DETR as the baselines. The results demonstrate that the Guidance Model had a positive impact, leading to a 4.52% improvement in the mR_{all} metric for VSL-Net and a 1.32% improvement for Moment-DETR in the same metric.

Takeaway. Our Guidance Model is general and can be combined with all grounding methods to bridge their performance from the short-form to the long-form setup. We refer the reader to the supplementary material for additional evaluations of the baselines on the short-form video grounding setup.

4.3. Ablation Study

In order to determine the optimal configuration of the Guidance Model, we conducted ablation studies focusing on three critical factors: **(i)** the selection of modalities (i.e., visual and/or audio), **(ii)** the performance of query-agnostic versus query-dependent guidance, and **(iii)** determining the optimal window size using the MAD dataset. Moreover, we performed an extensive investigation on actionless moments, which provides a significant differentiation of our implementation from the temporal proposals method. Moreover, we investigated the influence of the audio streams in video grounding models. Lastly, we conclude with a qualitative analysis of our proposed pipeline. Full detailed results in supplementary material.

Guidance through multimodality fusion. In this section, we investigate the contribution of having multiple modalities for the Guidance Model. In Table 3 we report the baselines' performance as-is (rows 1,5,9), as well as their performances when combined with our Guidance Model. The Guidance Model in each row was trained with different input modalities for the MAD dataset: (i) audio and text, (ii) video and text, (iii) audio, video, and text. We report Mean Recall@K (mR@K) for compactness.

For each baseline, the best performance can be achieved when all modalities are used for the Guidance Model, yielding an improvement between $1-18\times$ across all metrics. Using audio alone without visual input leads to modest performance improvement for Zero-shot CLIP, VLG-Net, and Moment-DETR. Additionally, visual clues can also boost all baselines, with Moment-DETR benefiting the most. Nonetheless, when combining all modalities, we can achieve the best overall performance.

Notice how the use of the Guidance Model is able to boost the performance of Moment-DETR, getting it to achieve a performance close to the proposal-based method even though the baseline performance is particularly poor. This finding allows us to conclude that strong short-video

Model	Modalities		mR@1	mR@5	mR@10	mR@50	mR@100
	Audio	Visual					
Zero-shot CLIP	✗	✗	3.7	10.0	13.9	27.3	35.0
	✓	✗	4.0	11.1	15.6	28.1	35.3
	✗	✓	4.8	12.0	16.1	29.2	36.0
	✓	✓	**5.2**	**12.6**	**16.7**	**29.7**	**36.4**
VLG-Net	✗	✗	2.5	8.6	13.4	31.0	40.8
	✓	✗	2.7	9.2	14.1	32.1	41.7
	✗	✓	3.5	11.1	16.7	35.0	44.1
	✓	✓	**3.9**	**12.1**	**17.8**	**36.0**	**45.2**
Moment-DETR	✗	✗	0.2	1.0	1.8	7.5	13.1
	✓	✗	1.0	4.2	7.1	20.7	29.5
	✗	✓	3.1	10.7	16.2	34.1	42.9
	✓	✓	**3.6**	**11.5**	**17.2**	**35.2**	**43.8**

Table 3: **Modality comparison for the Query Dependant Guidance Model.** The table reports the boost in performance achieved by the chosen baselines (rows 1,5,9) when combined with a Guidance Model having access to different input modalities. Mean Recall (mR@K) metric is computed over the validation set.

grounding methods can be bridged to long-form grounding through a two-stage approach that leverages a Guidance Model to reduce the search space for the temporal alignment of video and text.

Describable windows. The first ablation suggests that using both audio and video is a powerful representation. However, the Guidance Model used so far is query-dependent and leverages textual queries for identifying irrelevant windows. Our method can be very efficient if we can identify windows that are non-describable regardless of the input query so that the Guidance Model can process the video/audio streams only once. The question we want to answer in the following ablation is *"Can we devise an efficient first stage to identify non-describable windows and remove them from the search space of grounding methods?"* To investigate this research question, we devise a query-agnostic Guidance Model that does not process any textual query in input. In Table 4, we report the comparison between the query-agnostic and query-dependent Guidance Model, which uses both audio and visual cues as inputs. For compactness, we report the Mean Recall$_{all}$ (mR$_{all}$). Using audio-visual cues alone leads to improvements for Zero-shot CLIP, VLG-Net, and Moment-DETR. Conversely, when our Guidance Model takes the text query as input, it can better discriminate between relevant and irrelevant moments leading to consistent improvements across all baselines. We conclude that the query-dependent setup offers superior performance, at the cost of being more computationally expensive as the number of queries grows. On the other hand, the query-agnostic setup is computationally efficient as the video only has to be processed once by the Guidance Model, making it suitable for real-time or low-resource scenarios.

Model	Baseline	Query Agnostic		Query Dependent	
Zero-shot CLIP [33]	18.0	18.5	(+0.5)	**20.1**	(+2.1)
VLG-Net [34]	19.3	20.7	(+1.4)	**23.0**	(+3.7)
Moment-DETR [17]	4.7	8.3	(+3.6)	**22.3**	(+17.6)

Table 4: **Describable windows.** We report Mean Recall$_{all}$ for the baselines performances (second column), query agnostic guidance filter combined with baselines (third column), and query dependent guidance filter combined with baselines (fourth column). Performance is measured on the MAD validation set.

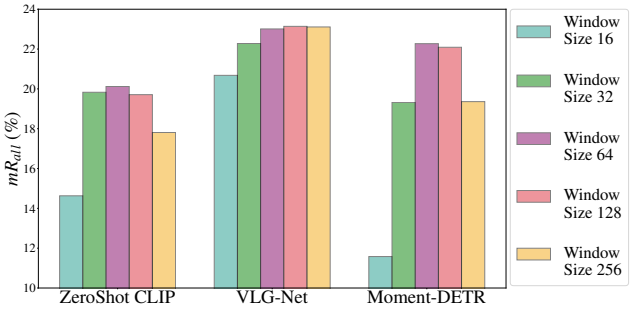

Figure 4: **Temporal field of view.** We report mR$_{all}$ for the combination of three baseline models (Zero-shot CLIP, VLG-Net, and Moment-DETR) with different *Guidance Models* trained with different window sizes. Results are computed on the validation split.

Temporal field of view. The next question we want to answer is: *"What is the optimal window size for the Guidance Model?"*. In Figure 4, we report the performance trend for all baselines when combined with a query-dependent audio-visual Guidance Model, where we vary the window size (temporal field of view) the Guidance Model can reason about. The trade-off in this experiment is between the fine-grained guidance versus the temporal context available to the guidance model. We test window sizes ranging from 16 to 256 time steps in the MAD dataset, with each time step accounting for 0.2s when the video FPS is set to 5. This yields models with a temporal field of view ranging from 3.2 to 51.2 seconds. For this experiment, we report the mR$_{all}$ metric.

Figure 4 showcases that, although short windows provide a more fine-grained guidance, they do not yield the best performance. For all methods investigated, the performance consistently increases from windows size 16 to 64. After such a window size, performance starts to decrease for the Zero-Shot CLIP and Moment-DETR, while little change is observed for VLG-Net. Following these findings, we fix the window size for the guidance model to be 64 in all other experiments. Hence, we can conclude that to improve the performance of the video grounding task, it is necessary to have sufficient context information so as to delimit all possible predictions correctly.

Method	mR@1	mR@5	mR@10
Zero-shot CLIP [33]	3.09	8.10	11.30
VLG-Net [34]	1.53	5.77	9.33
Moment-DETR [17]	0.15	0.65	0.91
†Zero-shot CLIP	4.31	10.38	14.01
†VLG-Net	2.32	8.07	13.13
† Moment-DETR	2.86	8.44	13.51

Table 5: **Describable windows beyond actions.** We compare the performance of our baseline models with the Guidance Model (†) and without guidance on actionless queries only. For this purpose, we extract actionless queries from MAD [33] test set and evaluate using mR@K for $K \in \{1, 5, 10\}$. The results showcase that our guidance method is not solely learning action-based concepts.

Describable windows beyond actions. Our motivation is to follow a data-driven approach to learn about what segment of the videos are worth describing. Looking at the data in MAD [33] reveals that a considerable portion of the queries (10% and 18%, respectively) lack verbs and describe only the environment and its attributes (adjectives, nouns). Examples of such queries include *"Night time at SOMEONE's building"* and *"Later, on a circular staircase"*. We recognize that interpreting this type of query is a challenging task, and it sets our approach apart from other computer vision methods, such as temporal proposals [4, 46, 41, 13, 28] because these moments do not contain any action element. Therefore, to validate the above hypothesis, we report in Table 5 the boost in performance that our Guidance Model brings for actionless queries. The metrics show an improvement when our method is applied to actionless queries, providing evidence for the effectiveness of our approach in capturing descriptions of actionless moments.

Enhancing Video Grounding by Using Audio. By incorporating audio features in Moment-DETR [17] using the MAD Dataset [33], we achieve a substantial performance boost. As shown in Table 6, the mRall (mean Recall over all classes) improves from 5.08% to 7.70%. Furthermore, when combined with the Guidance Model, the mRall further increases from 24.19% to 28.30%, demonstrating a consistent impact across setups. These results show a remarkable performance improvement when audio is integrated into the video grounding model.

We emphasize that our main focus is not on adding audio to the video grounding model but rather on the design and experimentation of the Guidance model. For a fair comparison, models in Tables 1 and 2 solely use visual and language features, excluding audio. However, we believe these results show encouraging results for the future inclusion of audio features in mainstream grounding models.

Method	Audio	m@R1	m@R5	m@R10	m@R50	m@R100	mR$_{all}$
Moment-DETR [17]	✗	0.24	1.11	2.02	7.92	14.13	5.08
Moment-DETR [17]	✓	0.70	2.15	5.11	11.23	19.08	7.70
†Moment-DETR [17]	✗	3.76	12.27	18.78	38.48	47.65	24.19
†Moment-DETR [17]	✓	**5.18**	**15.70**	**22.04**	**44.26**	**54.32**	**28.30**

Table 6: **Including audio in Grounding Model.** The results demonstrate a performance boost when audio is integrated into the video grounding model, thereby opening new possibilities for future research to explore this modality further. We enhance reader understanding by emphasizing methods that include the Guidance Model with (†) in their name.

Qualitative Results. Figure 5 presents three qualitative grounding results from the MAD [33] test set. The figure showcases how our Guidance Model can improve the ranking of moments predicted by a grounding model (*i.e.*, VLG-Net [34]). We report three positive cases (a-c) and a failure case (d). In Figure 5a, the Guidance Model is able to improve the ranking of the baseline model's prediction by 12 positions (from rank 16 to rank 4). Given the tight tIoU between the predicted moment and the ground truth, the Guidance Model is able to positively boost the performance. An even larger improvement is showcased in 5b, where the Guidance Model pushes the baseline prediction from rank 24 to 6. Figure 5c depicts a case when our Guidance Model helps to filter out a non-describable window by sensibly reducing its ranking from 5 to 62. Here, the moment predicted by the baseline model (blue) is ranked high, however, the predicted moment possesses the characteristics of a non-describable window. Therefore, our Guidance Model penalizes such a moment by reducing its ranking, making this false positive less relevant. Combined with our experimental section, the visualization provides a clearer understanding of the function of our guidance filter. Although the Guidance Model has an overall beneficial impact over all baselines, it can also impair performance in some cases. Figure 5d is such an example, where our Guidance Model worsens the ranking of a positive prediction; the reader should notice how this example depicts a static scene containing a moment relevant to the audience, with the Guidance Model penalizing it in a wrong way.

5. Limitations

Our method presents a notable improvement in video grounding performance. Nevertheless, it also presents a significant limitation regarding to the substantial inference time needed to match each query with every segment in a provided video. In our pursuit of finding a middle ground between computational efficiency and performance, we offer a more cost-effective option by introducing a query-agnostic framework. It is important to recognize that opting

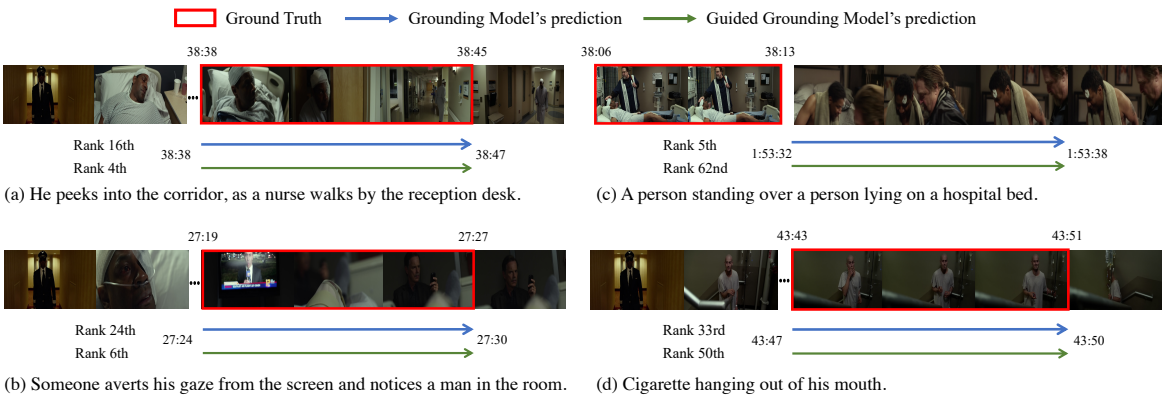

Figure 5: **Qualitative Results.** We compare ground truth annotation (red box) and the predicted temporal endpoints (arrows) with their respective rankings. We highlight in blue the prediction from the baseline VLG-Net [34] and in green the prediction of the baseline when combined with our guidance model. Notably, examples (a-c) depict a positive impact of the guidance model over the baseline one by either improving the ranking of a prediction or worsening it. Example (d) instead depicts a failure case for our pipeline.

for this alternative comes with a drawback: a decrease in performance across various metrics. This trade-off arises because the approach shift from query-dependent framework to detecting potentially interesting moments without considering a natural language query. Nonetheless, we believe this work will inspire the community in pursuing further effective and efficient video grounding on long-form video datasets.

6. Social impact

Our models development heavily leveraged the MAD dataset [33]. This dataset, primarily drawn from cinematic sources, underpins our training efforts but also introduces the biases inherent in movie portrayals. While providing benefits for training, this method also brings in biases come from historical stories, cultural portrayals, and societal norms seen in movies. These biases could affect the model's comprehension of real-world concepts. To address this, we conducted experiments using the Ego4D dataset [12]. This dataset takes a different approach by collecting videos worldwide in partnership with local organizations. Its goal is to capture a variety of everyday experiences across different cultures and areas, in order to offset any biases that may arise from how these experiences are depicted in movies. Impressively, our models demonstrated robust performance across both MAD and Ego4D datasets, showcasing their versatility and suggesting their potential for knowledge transfer. Moreover, the incorporation of Ego4D experiments contribute to the broader discourse on fair AI by promoting testing video-based approaches on such diverse datasets.

7. Conclusion

We presented a novel approach for language grounding in long videos. Our guidance framework is flexible, and our experimental section provides ample demonstration that it can boost performance for a variety of baselines, including proposal-based models like VLG-Net and zero-shot CLIP, as well as VSL-Net, and the proposal-free Moment-DETR. Our proposed Guidance Model can operate over different modalities and is very efficient in its query-agnostic version. However, we found that using the queries as part of the guidance mechanism allows for better performance. Future work includes exploring the use of smaller, more specialized Guidance Models to optimize inference time by prioritizing recall or precision based on application needs, effectively reducing inference time while maintaining desired performance levels. This could lead to a customized processing pipeline for various applications, achieving a balance between computational efficiency and accuracy. We also believe the community will further pursue two-stage approaches for the task at hand by designing more sophisticated and powerful guidance models that can provide even better assistance to paired grounding methods.

Acknowledgments. *W. Barrios* was supported by Dartmouth Graduate Fellowship through Guarini Office and Computer Science Department at Dartmouth College.

M. Soldan was supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research through the Visual Computing Center (VCC) funding, as well as, the SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence (SDAIA-KAUST AI).

A. Ceballos Arroyo acknowledges financial support for this work by the Fulbright U.S. Student Program, which is sponsored by the U.S. Department of State and Fulbright Colombia. In addition, he was partially supported by Northeastern University through a Graduate Fellowship.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing Moments in Video With Natural Language. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[2] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 609–617, 2017.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[4] Shyamal Buch, Victor Escorcia, Chuanqi Shen, Bernard Ghanem, and Juan Carlos Niebles. Sst: Single-stream temporal action proposals. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6373–6382, 2017.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[6] Chen Shaoxiang, Jiang Yu-Gang. Hierarchical Visual-Textual Graph for Temporal Activity Localization via Language. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[7] Jason Cramer, Ho-Hsiang Wu, Justin Salamon, and Juan Pablo Bello. Look, listen, and learn more: Design choices for deep audio embeddings. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3852–3856, 2019.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[10] Victor Escorcia, Mattia Soldan, Josef Sivic, Bernard Ghanem, and Bryan C. Russell. Temporal localization of moments in video collections with natural language. *CoRR*, abs/1907.12763, 2019.

[11] Gao Jiyang, Sun Chen, Yang Zhenheng, Nevatia, Ram. TALL: Temporal Activity Localization via Language Query. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[12] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Q. Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh K. Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Z. Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Christian Fuegen, Abrham Gebreselasie, Cristina González, James M. Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Yu Heng Khoo, Jáchym Kolár, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran K. Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Yunyi Zhu, Pablo Arbeláez, David J. Crandall, Dima Damen, Giovanni Maria Farinella, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard A. Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18973–18990, 2021.

[13] Tingting Han, Sicheng Zhao, Xiaoshuai Sun, and Jun Yu. Modeling long-term video semantic distribution for temporal action proposal generation. *Neurocomputing*, 490:217–225, 2022.

[14] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *ArXiv*, abs/1207.0580, 2012.

[15] Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, and Nicolas Carion. Mdetr - modulated detection for end-to-end multi-modal understanding. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1760–1770, 2021.

[16] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-Captioning Events in Videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[17] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 11846–11858. Curran Associates, Inc., 2021.

[18] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7327–7337, 2021.

[19] Kun Li, Dan Guo, and Meng Wang. Proposal-free video grounding with contextual pyramid network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(3):1902–1910, May 2021.

[20] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+language omni-representation pre-training. *ArXiv*, abs/2005.00200, 2020.

[21] Yang Li, Si Si, Gang Li, Cho-Jui Hsieh, and Samy Bengio. Learnable fourier features for multi-dimensional spatial positional encoding. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[22] Yanyu Li, Geng Yuan, Yang Wen, Eric Hu, Georgios Evangelidis, Sergey Tulyakov, Yanzhi Wang, and Jian Ren. Efficientformer: Vision transformers at mobilenet speed. *arXiv preprint arXiv:2206.01191*, 2022.

[23] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z XU, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022.

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[25] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021.

[26] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-Global Video-Text Interactions for Temporal Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[27] Abhishek Niranjan, Mukesh Sharma, Sai Bharath Chandra Gutha, and M Shaik. End-to-end whisper to natural speech conversion using modified transformer network. *arXiv preprint arXiv:2004.09347*, 2020.

[28] Zhiwu Qing, Haisheng Su, Weihao Gan, Dongliang Wang, Wei Wu, Xiang Wang, Yu Qiao, Junjie Yan, Changxin Gao, and Nong Sang. Temporal context aggregation network for temporal action proposal refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 485–494, 2021.

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[30] Merey Ramazanova, Victor Escorcia, Fabian Caba Heilbron, Chen Zhao, and Bernard Ghanem. Owl (observe, watch, listen): Localizing actions in egocentric video via audiovisual temporal context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2023.

[31] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding Action Descriptions in Videos. *Transactions of the Association for Computational Linguistics (ACL)*, 2013.

[32] Rodriguez Cristian, Marrese-Taylor Edison, Saleh Fatemeh Sadat, Li Hongdong, Gould Stephen. Proposal-free Temporal Moment Localization of a Natural-Language Query in Video using Guided Attention. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020.

[33] Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba, Chen Zhao, Silvio Giancola, and Bernard Ghanem. Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5026–5035, June 2022.

[34] Mattia Soldan, Mengmeng Xu, Sisi Qu, Jesper Tegner, and Bernard Ghanem. Vlg-net: Video-language graph matching network for video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3224–3234, 2021.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[37] Prateek Verma and Jonathan Berger. Audio transformers: Transformer architectures for large scale audio understanding. adieu convolutions. *arXiv preprint arXiv:2105.00335*, 2021.

[38] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020.

[39] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021.

[40] Mengmeng Xu, Mattia Soldan, Jialin Gao, Shuming Liu, Juan-Manuel Pérez-Rúa, and Bernard Ghanem. Boundary-denoising for video activity localization. *arXiv preprint arXiv:2304.02934*, 2023.

[41] Haosen Yang, Wenhao Wu, Lining Wang, Sheng Jin, Boyang Xia, Hongxun Yao, and Hujie Huang. Temporal action proposal generation with background constraint. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 3054–3062. AAAI Press, 2022.

[42] Zhuyu Yao, Jiangbo Ai, Boxun Li, and Chi Zhang. Efficient detr: Improving end-to-end object detector with dense prior. *ArXiv*, abs/2104.01318, 2021.

[43] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense Regression Network for Video Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[44] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6543–6554, Online, July 2020. Association for Computational Linguistics.

[45] Zhang Songyang, Peng Houwen, Fu Jianlong, Luo, Jiebo. Learning 2D Temporal Adjacent Networks for Moment Localization with Natural Language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[46] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *ICCV*, 2017.