Beyond Caption-Based Queries for Video Moment Retrieval

David Pujol-Perich[γ,δ*] Albert Clapés[γ,δ] Dima Damen[ζ] Sergio Escalera[γ,δ] Michael Wray[ζ]
[γ]University of Barcelona [δ]Computer Vision Center [ζ]University of Bristol
david.pujolperich@ub.edu

Abstract

Current Video Moment Retrieval (VMR) models are trained on videos paired with captions, which are written by annotators after watching the videos. These captions are used as textual queries—which we term ***caption-based queries****. This annotation process induces a visual bias, leading to overly descriptive and fine-grained queries, which significantly differ from the more general* ***search queries*** *that users are likely to employ in practice.*

In this work, we investigate the degradation of existing VMR methods, particularly of DETR architectures, when trained on caption-based queries but evaluated on search queries. For this, we introduce three benchmarks by modifying the textual queries in three public VMR datasets—i.e., HD-EPIC, YouCook2 and ActivityNet-Captions. Our analysis reveals two key generalization challenges: (i) A language gap, arising from the linguistic under-specification of search queries, and (ii) a multi-moment gap, caused by the shift from single-moment to multi-moment queries. We also identify a critical issue in these architectures—an active decoder-query collapse—as a primary cause of the poor generalization to multi-moment instances. We mitigate this issue with architectural modifications that effectively increase the number of active decoder queries. Extensive experiments demonstrate that our approach improves performance on search queries by up to 14.82% mAP_m*, and up to* 21.83% mAP_m *on multi-moment search queries. The code, models and data are available in the project webpage.*

1. Introduction

Video Moment Retrieval (VMR) aims to localize temporal segments in a video given a user-defined textual query. While current models achieve remarkable success on existing benchmarks, in this work we raise awareness of a key limitation of VMR datasets: text queries are defined using the annotated captions, written after annotators watch the videos. These captions, which we name *caption-based queries*, induce a visual bias—overly descriptive visually-informed textual annotations. In contrast, real users interact through *search queries*, often formulated without watching the video, which can be of a more general and *under-specified* nature [31]. For instance, while an annotator might formulate a caption-based query *"a man in a yellow jersey intercepts a loose pass from the opposing team near the box and scores a powerful volley"*, a typical search query can be *"when are goals being scored?"* (see Fig. 1).

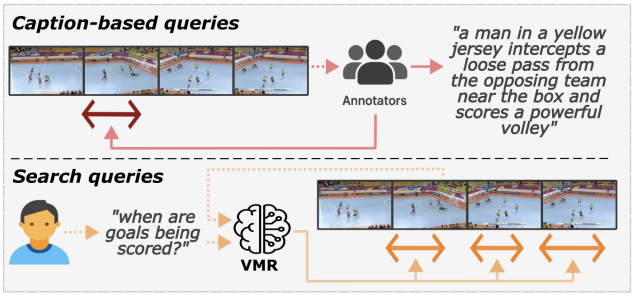

Figure 1. After watching a video, annotators write detailed, visually-informed captions that map to a single GT moment. However, at inference time, users formulate less detailed, visually-uninformed search queries that often map to multiple GT moments.

Devising VMR methods that robustly perform on search queries requires suitable benchmarks, different from the existing caption-based ones [15, 16, 27, 28]. However, collecting new search-query datasets remains an open challenge, as it is unclear how text annotation and video observation can be decoupled in a feasible manner. We instead repurpose existing datasets—which include paired videos and captions—by proposing a pipeline that under-specifies captions. We systematically explore levels of under-specificity, by changing the level of details available in the original caption. We consequently propose three $\{S\}$earch-query benchmarks: *HD-EPIC-S*$\{1,2,3\}$, *YC2-S* and *ANC-S* based on HD-EPIC, YouCook2 and ActivityNet-Captions.

DETR [5] has become the cornerstone of most existing VMR methods [14, 24, 33, 44], thanks to its usage of K learnable decoder queries, each of which maps to a potential retrieved moment with a corresponding confidence score. Our evaluation indicates that these methods, trained on caption-based queries, substantially degrade when evaluated on more under-specified search queries. We iden-

*Work partially completed whilst at University of Bristol.

tify two key factors driving this degradation: (i) a *language gap*, reflecting the linguistic distribution shift between caption and search queries, and (ii) a *multi-moment gap*, arising from how caption-based queries map to a single ground-truth (GT) moment, while the under-specified search queries often map to multiple moments.

In this paper, we quantify the impact of both the language and multi-moment gaps, and particularly address the former by proposing architecture modifications including the removal of self-attention mechanisms and the addition of a decoder-query dropout regularizer. These modifications improve generalization to search queries without the expensive task of re-annotating VMR training sets.

In short, our contributions are: 1) we explore the task of VMR beyond using captions as textual queries. We reformulate these as under-specified versions of existing captions, so they are closer to common user-defined queries, while still making the most of available benchmarks; 2) we create three VMR benchmarks with search queries by mapping caption-based queries to under-specified search queries; 3) we demonstrate the significant degradation in performance and identify its two main causes: a *language* and a *multi-moment gap*; and 4) we mitigate this degradation, particularly that induced by the multi-moment gap, by introducing targeted architectural modifications that boost generalization to search queries.

2. Related work

Video Moment Retrieval has become a cornerstone in video understanding, aiming to localize start-end times of moments in a video, based on textual queries. Existing approaches can be broadly categorized into *proposal-based* and *proposal-free* methods. *Proposal-based* approaches generate candidate temporal segments through temporal anchors [3, 11, 37] or sliding-windows [1, 6]. However, their performance heavily depends on the quality and redundancy of these proposals. In contrast, *proposal-free* methods avoid the explicit candidate generation, leveraging a single-stage architecture to predict the moment segments. Most of these works adopt DETR-based architectures [5], which refine a fixed set of learnable decoder queries, each of which represents a candidate segment. This paradigm was introduced by [16], with follow-ups like [24, 25, 44] improving various aspects such as the cross-modality modules, recursive decoding schemes, etc. Our work also focuses on the DETR architecture, as this is the foundation of the majority of existing state-of-the-art VMR methods [24, 33, 34, 43, 44], thus maximizing the impact of our findings.

Generalization of VMR methods: While DETR-based VMR models have shown remarkable success in existing VMR benchmarks, their generalization capabilities beyond their training data distribution remains largely underexplored. Most existing works address this from a vision-centric perspective, analyzing aspects like action duration and temporal shifts [21, 26, 42] or temporal biases [18, 26, 40]. More recent studies have begun exploring the role of language biases in generalization [19, 22], tackling rare-word usage and grammatical mistakes [39] and the use of unlabeled data to improve language robustness [2]. In parallel, research from the multi-modal and linguistic community [29, 32] emphasizes how under-specification and contextual variation also impact the generalization capabilities of models. Moreover, Liang *et al*. [20] address the effect of imprecise queries by incorporating them into training for the task of ranked retrieval across corpora. Instead, we tackle the alternative challenge of multi-moment retrieval within a single video by generalizing to such queries in a zero-shot manner. We find that a fundamental bottleneck in VMR stems from the use of visually-informed captions as training queries. This induces a visual bias toward overly descriptive language, often misaligned with more abstract, visually-uninformed queries prompted by users in real-world scenarios.

Query collapse in DETR: A core issue that we address is the query collapse in DETR architectures, whereby only a small subset of decoder queries meaningfully contribute to the final prediction(s), while the rest remain inactive. This issue, reported in object detection [17, 23, 47], temporal action detection [14], and 3D detection [38, 46], is largely attributed to the sparse supervision from the one-to-one matching [5]. We observe a similar phenomenon in VMR, driven instead by the single-moment prior of existing benchmarks, which typically provide one annotated moment per query. This leads to a significant query collapse that hinders generalization to multi-moment queries. While some works introduce alternative mechanisms to provide additional supervision signals [7, 10, 13, 17, 36, 43], these mainly target accelerating convergence, proving unable to overcome this strong prior. Curating new datasets with multiple annotated moments per query could alleviate this issue [4, 16], but would entail costly re-annotations or directly discarding most existing datasets. This motivates our approach, which introduces architectural modifications that counter the single-moment prior, leveraging existing datasets while improving generalization to unseen multi-moment scenarios.

3. Problem definition and benchmarking

3.1. Problem definition

Video Moment Retrieval (VMR) is defined as follows: Given a video–query pair (v_i, q_i), the task is to predict the start–end times $\{(s_j, e_j)\}_{j=1}^{M_i}$ of all temporal segments in v_i corresponding to the textual query q_i—i.e. moments. In this work we revisit this task, and focus on the often overlooked aspect of how textual queries are defined.

Specifically, we highlight the underlying assumption

of all existing VMR benchmarks where queries are created from the captioning annotations—i.e. annotators who watch the videos before writing a sentence that best describes the moment [11, 15, 28, 45]. These *caption-based queries* induce a visual bias that the queries perfectly match the description of the moment, thus descriptive and fine-grained in nature. In contrast, real-world users formulating their textual queries are normally unaware of the detailed content of the video, relying instead on broader, *under-specified* descriptions. Such queries can range from moderately detailed to very general ones, differing substantially from captions. We refer to these as *search queries*.

To model the distribution shift between caption and search queries, we derive $\mathcal{Q}_{search}$ from $\mathcal{Q}_{caption}$ through varying degrees of under-specification. We thus capture the shift from visually-informed to visually-uninformed textual queries. This allows us to study how VMR methods trained on caption-based queries $\mathcal{Q}_{caption}$ perform on search queries $\mathcal{Q}_{search}$, more closely aligned with common situations in which users either do not know or do not exactly remember the contents of the video.

3.2. Benchmarks

Common VMR benchmarks rely on caption-based queries, and thus do not include search-query annotations. Re-annotating benchmarks is a challenging task, as it is unclear how to disentangle search-query annotation from video observation in a feasible and scalable manner. Accordingly, we propose to make the most of existing datasets, introducing a pipeline that rewrites a densely-annotated caption-based dataset, to a search-query variant. Dense temporal annotations are essential here since, as queries become more under-specified, these may correspond to multiple additional moments in the video. If dense annotations are provided, the search for these new correspondences can be done automatically. This contrasts with sparsely annotated datasets [16], which would require extensive manual re-annotation to cover unlabeled moments corresponding to the under-specified search query. In particular, we use three densely-annotated public datasets to introduce our search-query pipelines. We detail our pipeline next.

3.2.1. Search-query pipeline

Our pipeline (see Fig. 2) comprises two main stages:

1) **Per-query under-specification stage:** To simulate the potential ambiguity of search queries, we generate under-specified queries from fine-grained ones through an LLM-based pipeline. Specifically, we instantiate two cooperative agents based on Gemma-12B [35]: a rewriter and a validator. The rewriter agent receives a fine-grained caption and rewrites it into a less detailed version while preserving the core semantics. For example, the caption *"a man tying his running shoes before starting a marathon"*, will be rewritten as *"a person getting ready to exercise"*. This step allows us to model alternative under-specified search queries where users may omit context information like subject, object or intent. Inspired by [8], we prevent hallucinations by introducing a second agent that acts as a validator. This agent flags any inconsistent rewritings, which are subsequently corrected by human annotators.

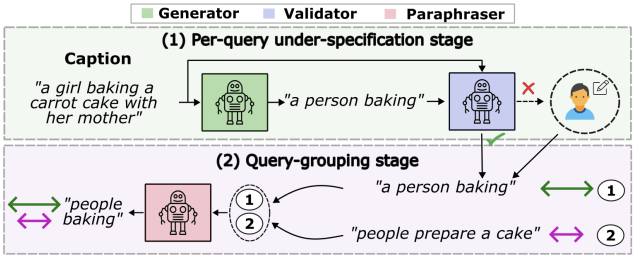

Figure 2. Overview of the search-query pipeline. Each of the caption is first processed by an agent that generates per-query under-specifications, which are validated by a second identical agent and manually re-annotated if abnormal. Individual queries mapping to the same under-specified query are then grouped, and a final agent produces a representative search query per group.

2) **Query-grouping stage:** When a query is under-specified, it can correspond to multiple valid moments in the video, since several fine-grained situations can fit the same broader description. For example, *"a person cooking food"* could match segments showing *"a man sweating onions"*, or *"a woman stirring soup"*. This makes it essential to group all original fine-grained queries that map to the same or highly similar under-specified query. To perform this grouping, we compute pairwise similarities across all under-specified queries using a pre-trained transformer-based sentence encoder [30]. Queries with a high similarity are merged into the same group, forming a multi-moment instance. We then use an LLM-based aggregator that summarizes each group into a single representative under-specified query, removing minor differences across group members while keeping their shared semantics. Find more details, including prompts, in Sec. C.1.

3.2.2. Search-based VMR benchmarks

The proposed search-query pipeline enables us to introduce three search-query benchmarks, denoted by "-S":

HD-EPIC-S{1,2,3}: HD-EPIC [28] is a large-scale egocentric dataset featuring long cooking videos. Due to the exceptional level of detail of its annotated captions (16.47 words per query on avg.), we derive three progressively under-specified variants—i.e., S1, S2, S3—by gradually removing contextual details. For example, *"Pick up a tissue from inside the plate on the countertop using the right hand"* → *"Pick up a tissue from the plate"* (S1) → *"Pick up a tissue"* (S2) → *"Pick up something"* (S3).

YouCook2 (YC2)-S: YC2 [27] contains step-level narrations for instructional cooking videos. YC2-S replaces the original detailed descriptions with under-specified versions,

Table 1. Statistics of the search-based VMR benchmarks

Dataset	# videos	# queries	Duration per video(s)	Moments per query	Moments per query (multi)	# multi. queries	% multi. queries	Query length
HD_EPIC [28]	156	59,454	954	1.00	0.00	0	0.00	16.47 ± 7.9
HD_EPIC-S1	156	36,819	954	1.61	3.64	8,560	23.25	6.09 ± 2.5
HD_EPIC-S2	156	31,521	954	1.88	3.87	9,717	30.83	3.73 ± 1.2
HD_EPIC-S3	156	10,266	954	5.79	11.09	4,873	47.47	3.031.0
ANC [15]	14950	71,957	152.8	1.00	0.00	0	0.00	13.16 ± 6.1
ANC-S	14950	59,138	152.8	1.21	2.45	8,818	14.91	4.83 ± 1.4
YC2 [27]	2268	13,829	326	1.00	0.00	0	0.00	8.86 ± 3.97
YC2-S	2268	7,466	326	1.84	2.97	3,212	43.02	2.08 ± 0.50

mostly emphasizing the main actions. For example *"Add salt to the pan and mix"* $\rightarrow$ *"Season food"*.

ActivityNet-Captions (ANC)-S: ActivityNet-Captions [15] consists of third-person open-domain videos. ANC-S derives under-specified versions of its corresponding textual queries, removing fine-grained contextual details. For example, *"The man in white shirt is strumming the bongo drum."* $\rightarrow$ *"A person plays an instrument"*. Unlike previous benchmarks, ANC-S presents a considerable number of multi-moment search queries with overlapping moments.

Tab. 1 depicts the main statistics of these benchmarks. Using our pipeline, we extend existing single-moment datasets into new multi-moment versions, where up to 47.57% of the queries correspond to multiple moments. Due to the linguistic under-specification, the average query length is also reduced by up to 82%. All this while ensuring the realism and reliability of our generated search queries, validated in Sec. E and Sec. J.

3.3. Metrics for Search-Based VMR

VMR performance is typically measured using Recall@1 (R1) and mean Average Precision (mAP), capturing top-1 accuracy and overall ranking quality, respectively. However, these metrics are inadequate when evaluating multi-moment queries. When retrieving under-specified queries these often map to additional GT moments. For example, while caption-based queries might target *"a man in a black shirt enters through the kitchen door"* retrieving a single moment, a more general search query *"a person entering a room"* could naturally map to multiple moments, including the previous. This setting thus requires metrics that estimate how individual moments are retrieved, regardless of whether they arise alone or alongside other moments.

Existing metrics like R1 or mAP are unsuitable for two reasons: First, recall metrics like R1 only evaluate accuracy over the top-k predictions—i.e., $k = 1$ for R1. While appropriate when queries map to exactly k moments, these metrics provide an incomplete evaluation when queries map to more than k moments. For instance, for a 2-moment query, R1 assesses if the top-1 prediction matches *any* GT, ignoring if the other GT was retrieved at all. Second, metrics like mAP aggregate all GT moments of a query into a single video-query score, obscuring per-moment retrieval quality. Consider a query q_1 that maps to a GT moment g_1. If a model fails to retrieve it, mAP would clearly indicate failure. However, for a more general query q_2 that maps 4 moments (g_1-g_4), the same model might detect g_2-g_4 but still miss g_1. In this case, mAP would remain high, masking the error of g_1. Thus, the retrieval quality of an individual moment depends on how many moments co-occur with it, making it unsuitable for a fair evaluation.

We overcome these issues by introducing their respective multi-moment extensions, denoted as R_m and mAP_m:

Multi-moment recall (R_m): R_m generalizes the R1 metric to handle textual queries associated with multiple GT moments. Concretely, instead of considering the entire query as correct if one of the GT moments is retrieved with the highest confidence, R_m evaluates each GT moment independently, checking if at least one of the top predictions correctly matches it. More specifically, we consider a given moment g_i to be correctly retrieved under a certain IOU threshold τ ($R_m(g_i, \tau) = 1$), if 1) the prediction detecting it had the highest confidence, or 2) all the predictions with higher confidences successfully retrieved other GT moments. The latter avoids the predictions that retrieved other valid GT moments—as there may be multiple ones—penalize the retrieval quality of g_i.

Averaging across all GT moments $\mathcal{G}$ of the dataset yields:

$$R_m(\tau) = \frac{1}{|\mathcal{G}|} \sum_{g_i \in \mathcal{G}} R_m(g_i, \tau) \tag{1}$$

Intuitively, R_m measures whether a model retrieves each GT moment with a top-confidence prediction, without interference from the other GT moments in the video.

Multi-moment mAP (mAP_m): To make mAP sensitive to multiple GT, mAP_m evaluates each GT individually. For a given g_i and threshold τ: (1) predictions intersecting g_i with IOU$\geq \tau$ are considered true positives, (2) predictions not matching any GT are considered false positives, and (3) predictions matching other GT moments (not g_i) are ignored to avoid penalizing g_i for a different, yet correct, prediction.

With this, we compute the precision-recall curve (P_i, R_i) for each g_i, and apply the standard mAP interpolation [9] to obtain its individual score $AP_m(g_i, \tau)$. We then average across these per-GT scores for a given threshold τ:

$$mAP_m(\tau) = \frac{1}{|\mathcal{G}|} \sum_{g_i \in \mathcal{G}} AP_m(g_i, \tau), \tag{2}$$

Intuitively, mAP_m measures how well a model retrieves each of the GT moments, without interference from the other GT moments. See Sec. A for more details.

4. Search-Based VMR

Here, we analyze how current VMR methods trained on caption-based datasets perform when evaluated on search queries, and how to mitigate the multi-moment gap.

4.1. Evaluating caption-based models

In our experiments, we evaluate two representative models—i.e., CG-DETR [24] and LD-DETR [44]—on the three proposed benchmarks. For each, we train on the training set of the corresponding caption-based dataset and evaluate on the same test set for both the original caption-based dataset and our proposed search-query benchmark—e.g., train on the training set of HD-EPIC, and evaluate on the test set of HD-EPIC and HD-EPIC-S2. For ANC-S and YC2-S we leverage the original splits from [15, 41], while for HD-EPIC we create an 80-20 train/test split.

Figure 3 (left) shows the progressive performance decay across HD-EPIC-S1/S2/S3 benchmarks, with a relative degradation of up to 71.75% and 77.40% of R_m@0.3 for CG-DETR and LD-DETR, respectively. Similar trends are observed on ANC-S (center) and YC2-S (right) with drops of up to 31.8% and 60.7% of R_m@0.3, respectively. These drops evidence a substantial shift between caption-based queries and search queries, showing that existing models, when trained solely on descriptive caption-based queries, significantly degrade on less detailed search queries. This inevitably hinders the deployment of existing VMR systems in real-life scenarios (see Sec. 5.2 for the full results).

Below, we isolate two main causes of this degradation:

- Language gap: The linguistic shift between caption and search queries, characterized by missing visual details, looser references or the use of more abstract nouns—e.g., *"food"* instead of *"green peppers"*.
- Multi-moment gap: The gap arising from the mismatch between training on single-moment queries and evaluating on multi-moment ones, as under-specified search queries often correspond to multiple moments.

To quantify the effect of these factors, we partition the test set of the search queries into two subsets: $\mathcal{D}^{search}_{single}$ containing under-specified queries that map to a single GT moment, and $\mathcal{D}^{search}_{multi}$ containing search-queries that map to multiple GT moments.

For a fair one-to-one comparison across specificity levels, we also partition the original caption-based dataset $\mathcal{D}^{caption}$ using the same moment correspondence. Thus, $\mathcal{D}^{caption}_{single}$ and $\mathcal{D}^{caption}_{multi}$ contain the same moments, as $\mathcal{D}^{search}_{single}$ and $\mathcal{D}^{search}_{multi}$, respectively, but paired with their more specific captions. This yields our evaluation setup:

$$(\mathcal{D}^{caption}_{single}, \mathcal{D}^{search}_{single},) \text{ and } (\mathcal{D}^{caption}_{multi}, \mathcal{D}^{search}_{multi},)$$

allowing us to measure degradation due to purely linguistic changes ("single" split) versus the compounded effect of linguistics and multi-moment mapping ("multi" split).

Figure 4 reports the performance of CG-DETR on the three benchmarks after the decoupling of "single" and "multi" instances. For HD-EPIC, the language gap (performance on single) increases as we progressively evaluate more under-specific search queries, dropping from 15.9% to 49.6% with respect to the original mAP_m@0.3. Importantly, the compounded effect of the language and multi-moment gaps (performance on "multi") aggravates this effect further, reaching a degradation on HD-EPIC-S3 of 73.8% compared to the original mAP_m@0.3. This highlights the significant additional impact on performance of the multi-moment gap. These observations remain consistent across all benchmarks.

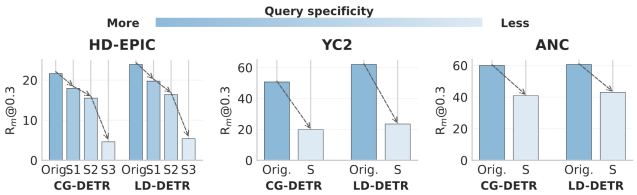

Figure 3. Evaluation of the representative models on both the original datasets and their corresponding search query extensions.

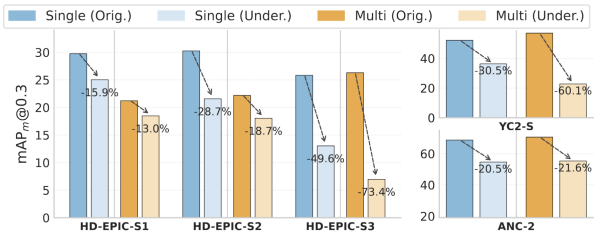

Figure 4. Performance degradation for CG-DETR on caption versus search-based evaluation for the "single" and "multi" splits.

We next focus on addressing this multi-moment gap, an aspect largely under-explored in the VMR literature and key to the model architecture design. We leave addressing the language gap for future work, as it may be resolved with more advanced vision-language models, capable of reasoning across varying levels of specificity.

4.2. Mitigating the multi-moment gap

In this section, we analyze the underlying causes of the performance degradation of VMR models on multi-moment query setups. We argue that this degradation mostly stems from misalignment between caption-based training data—characterized by a single relevant moment as GT—and search-based evaluation data, which frequently contains multiple valid moments. This discrepancy induces a strong single-moment prior—i.e., a bias towards expecting a single GT moment per video-query pair. One could attempt to resolve this by curating more diverse training data, which is impractical due to annotation costs and the uncertain feasibility of devising true search-query datasets. We instead approach this issue purely from a model's perspective which allows us to reuse all existing VMR training regimes.

4.2.1. Implications of a single-moment prior

We next analyze why DETR-based methods trained on single-moment queries struggle with multi-moment queries at inference time. Concretely, since each decoder query

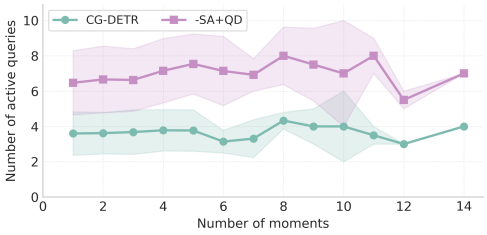

Figure 5. Visualization of the active query collapse on HD-EPIC-S2 for the base CG-DETR and our method -SA+QD.

produces a candidate moment, analyzing the number of active decoder queries—i.e., those whose confidence does not vanish—directly reflects the model's capacity to retrieve multiple moments. In VMR, the number of active decoder queries can also be thought as the "compute budget" available to retrieve all the GT moments. When the number of active queries does not scale with the number of moments of the target instance, the model cannot retrieve all moments—e.g., when only 2 queries are activated in a 4-moment instance, the upper bound of retrieved moments is 50%. We term this phenomenon *active decoder-query collapse*.

As shown in Fig. 5, this phenomenon indeed affects VMR methods. Concretely, VMR methods trained on standard caption-based datasets (blue line) yield an insufficient number of active decoder queries when evaluated on search queries—around 4, regardless of the number of moments of the search queries (x-axis). This stems from the single-moment training, which impedes generalization beyond queries mapping to 4 moments.

4.2.2. Addressing the active decoder-query collapse

Having identified the *active decoder-query collapse* as a key limitation for generalization to multi-moment queries, we explore whether this can be mitigated purely from a model perspective, without altering the standard VMR training regimes and datasets. We find that this is possible by preventing models from overfitting to the single-moment prior encoded in the training data. Specifically, we identify two structural causes of this prior fitting: (i) Coordination collapse, where the self-attention mechanism causes decoder queries to suppress one another, and (ii) Index collapse, where a fixed, small subset of decoder query indices dominate activation.

In the following, we introduce architectural modifications that retain the model's capacity to learn the VMR task, while mitigating these forms of collapse.

Coordination collapse: The first cause arises from how the self-attention (SA) within each decoder layer enforces coordination among decoder queries. This drives them to "agree" on which query should handle the GT moment and which should remain inactive.

Formally, a standard decoder layer can be defined as:

$$\hat{Q}^{l+1} = FFN(CA(SA(\hat{Q}^{l}), M) \tag{3}$$

where $M \in \mathbb{R}^{T \times F}$ are the fused multi-modal features, CA denotes cross-attention and FFN a feed-forward network.

As noted by [12], the CA module injects the cross-modality information, while the role of SA is pushing decoder queries apart from each other to avoid redundancy. However, unintentionally, this also drives the majority of decoder queries to deactivate.

Interestingly, we find that an effective way of overcoming this issue is removing this SA module altogether, while leaving the losses unchanged:

$$Q^{l+1} = FFN(CA(Q^{l}, M) \tag{4}$$

Eliminating this inter-query communication prevents these coordination-based shortcuts, encouraging each decoder query to act independently. However, this also removes the model's built-in mechanism for avoiding redundant predictions. We address this by applying Non-Maximal-Suppression (NMS) during post-processing, which filters out overlapping/redundant predictions.

Index collapse: Mitigating the coordination collapse alone is insufficient as the model is still able to overfit to the single-moment prior, and thus still suffers active decoder-query collapse. The reason resides in an *index collapse*, where the same decoder query indices repeatedly dominate the output confidence, while the rest remain inactive—e.g., decoder queries with index 1–4 are the only ones ever activating. During training, the single-moment prior drives the model to associate the detection of the single GT moment with only a handful of fixed decoder query indices based on their learnable initializations. This dominance is progressively reinforced throughout training, leaving the rest of the decoder queries permanently inactive and thus, unused.

We counter this effect by applying a targeted query dropout strategy, which randomly zeroes out $k\%$ of the learnable queries $Q \in \mathbb{R}^{Q \times F}$ during each training iteration:

$$\hat{Q} = Q \odot M, \quad M \sim \mathbb{B}(1-k) \tag{5}$$

where $\mathbb{B}$ is sampling from the Bernoulli distribution with keep probability $(1-k)$. This regularization promotes the model to distribute supervision across more queries, preventing over-reliance on a fixed subset.

Together, these modifications considerably reduce the number of permanently inactive indices, resulting in a consistent increase in the number of active decoder queries (see Fig. 5 orange line), boosting search-query generalization.

5. Experimentation

5.1. Experimental setup

The following experiments evaluate how baselines trained on a caption-based dataset generalize to search-based

Table 2. Results of both CG-DETR and LD-DETR on HD-EPIC-S{1,2,3} benchmarks with respect to our proposed modifications.

Model	Input	Variant	R_m @0.1	R_m @0.3	R_m @0.5	R_m Avg.	mAP_m @0.1	mAP_m @0.3	mAP_m @0.5	mAP_m Avg.
CG-DETR	S1	base	28.61	17.95	8.99	18.51	36.21	22.84	11.59	23.54
		-SA+QD	**29.87**	**19.69**	**10.86**	**20.14**	**39.74**	**26.49**	**14.87**	**27.03**
	S2	base	24.71	15.52	7.89	16.04	32.15	20.1	10.29	20.84
		-SA+QD	**26.17**	**17.00**	**9.40**	**17.52**	**35.38**	**23.39**	**13.04**	**23.93**
	S3	base	9.50	4.61	2.08	5.39	16.20	8.01	3.58	9.26
		-SA+QD	**10.57**	**6.52**	**3.45**	**6.84**	**17.27**	**10.65**	**5.54**	**11.15**
LD-DETR	S1	base	29.42	19.77	10.50	19.89	36.55	24.50	13.18	24.74
		-SA+QD	**30.18**	**20.26**	**10.83**	**20.42**	**40.5**	**27.54**	**14.94**	**27.66**
	S2	base	25.23	16.38	8.46	16.69	32.42	21.11	10.93	21.48
		-SA+QD	**26.36**	**16.98**	**8.87**	**17.40**	**36.37**	**23.75**	**12.54**	**24.22**
	S3	base	**10.44**	**5.37**	**2.58**	**6.13**	16.48	8.65	4.11	9.74
		-SA+QD	**10.44**	5.28	2.39	6.03	**17.79**	**9.06**	**4.19**	**10.34**

Table 3. Results of both CG-DETR and LD-DETR on YC2-S with respect to our proposed modification.

Model	Variant	R_m @0.1	R_m @0.3	R_m @0.5	R_m Avg.	mAP_m @0.1	mAP_m @0.3	mAP_m @0.5	mAP_m Avg.
CG-DETR	base	28.92	19.87	11.22	20.00	38.83	26.96	15.21	27.00
	-SA+QD	**29.97**	**20.32**	**11.38**	**20.55**	**41.00**	**29.40**	**17.21**	**29.20**
LD-DETR	base	33.13	23.48	11.70	22.70	41.69	30.04	15.58	29.10
	-SA+QD	**35.86**	**24.76**	**13.17**	**24.59**	**45.66**	**33.09**	**18.74**	**32.49**

Table 4. Results of both CG-DETR and LD-DETR on ANC-S with respect to our proposed modification.

Model	Variant	R_m @0.1	R_m @0.3	R_m @0.5	R_m Avg.	mAP_m @0.1	mAP_m @0.3	mAP_m @0.5	mAP_m Avg.
CG-DETR	base	60.44	40.89	24.56	41.96	72.12	54.92	36.42	54.48
	-SA+QD	**63.75**	**43.12**	**25.50**	**44.12**	**74.00**	**56.42**	**38.20**	**56.20**
LD-DETR	base	62.58	43.00	**26.08**	43.88	73.35	56.17	**36.79**	55.43
	-SA+QD	**65.21**	**43.89**	25.77	**44.95**	**74.25**	**56.31**	36.69	**55.75**

benchmarks. Specifically, we evaluate HD-EPIC-S, YC2-S and ANC-S; and report the R_m and mAP_m, on IOU $\{0.1, 0.3, 0.5\}$. Full implementation details are in Sec. C.

5.2. Main experiments

Do our proposed modifications bridge the generalization gap to search queries? From the results, we observe that our proposed architectural modifications, hereby denoted as (-SA+QD), substantially improve performance across all search-query datasets. For instance, on HD-EPIC-S2 (see Tab. 2) these modifications increase R_m@0.1 from 24.71 to 26.71 and the mAP_m@0.1 from 32.15 to 35.38. Similarly, on YC2-S (see Tab. 3) we observe an absolute improvement of up to 2.96 mAP_m@0.3. Moreover, even with the smaller multi-moment gap for ANC-S, we observe that our modifications lead to comparable or improved performance across all metrics (see Tab. 4). To contextualize these gains, we also compare against an oracle of the base model, which is trained directly on the under-specified queries—thus perfectly matching the evaluation specificity, unlike the models trained on captions. Our approach recovers nearly 70% of the oracle gap, confirming the effectiveness of (-SA+QD) in bridging the multi-moment gap, thus improving generalization to search queries. Full results can be found in Sec. D.

Where do these gains come from? To disentangle the benefits of our proposed modifications, we separately evaluate single and multi-moment instances. Figure 6 shows

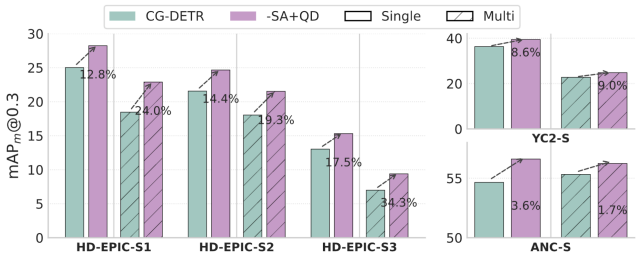

Figure 6. Dissection of the performance of CG-DETR on HD-EPIC-S2, for single, multi and all the instances, respectively.

that while performance of (-SA+QD) on single-moment queries improves modestly, in most cases there is a prominent improvement on multi-moment queries by up to 34.3% mAP_m@0.3. This confirms that our method specifically benefits multi-moment queries while also improving single-moment cases. See Sec. D. for the extended results.

5.3. Ablations

Below, we ablate key aspects of our findings, reporting results for CG-DETR on HD-EPIC-S2. We report the average R_m and mAP_m across IoU values $\{0.1, 0.3, 0.5\}$. Additional ablations can be found in Supp.

Alternative methods for decoder query activation: We examine whether alternative methods can increase the number of active queries and yield comparable gains to our proposal. Specifically, we evaluate two families of approaches (see Tab. 5): (i) alternative matching strategies that provide supervision to multiple decoder queries, and (ii) a data-augmentation scheme that simulates multi-moment setups.

While matching strategies increase the number of active queries, these activated queries produce redundant predictions—predicting nearly identical segments around the same GT moment. This can be observed in the number of predictions overlapping a GT (%match P), which nearly doubles, while the number of retrieved GT moments (%match GT) decreases, leading to a drop in generalization.

Similarly, data-augmentation techniques that replicate GT moments in different video locations also fail to improve generalization. We attribute this to the disruption of temporal coherence, which leads to overfitting.

Overall, these results confirm that merely increasing the number of active queries through additional supervision is insufficient; effective generalization to multi-moment setups also requires diversity-promoting mechanisms that encourage complementary behavior in decoder queries.

Preserving diversity via 1-to-1 matching: The previous ablation showed that merely activating more queries does not improve generalization if diversity is not preserved. Our proposal addresses this, increasing the number of active queries while also maintaining diversity among them. This is achieved by keeping the 1-to-1 matcher [5]. This strategy enforces competition between decoder queries, prevent-

Table 5. Ablation of methods to increase number of active queries.

Variant	R_m	mAP_m	# active	% match P	% match GT
base	16.04	20.84	3.64 ± 1.18	0.06 ± 0.07	0.36 ± 0.35
+ 1-to-5 matching [17]	14.66	16.30	9.56 ± 3.20	0.11 ± 0.16	0.21 ± 0.28
+ 1-to-k matching [17]	10.78	11.01	20.00 ± 0.0	0.16 ± 0.33	0.07 ± 0.18
+group_matching [7]	15.34	17.97	8.69 ± 3.08	0.10 ± 0.13	0.27 ± 0.31
+hybrid [13]	14.67	17.04	8.68 ± 2.90	0.10 ± 0.13	0.28 ± 0.32
+ms_matcher [43]	15.75	20.94	3.58 ± 1.14	0.06 ± 0.07	0.36 ± 0.34
+data_augmentation	13.85	21.37	4.68 ± 1.78	0.07 ± 0.08	0.38 ± 0.35
-SA+QD (ours)	**17.52**	**23.93**	6.43 ± 2.16	0.11 ± 0.13	0.42 ± 0.37

Table 6. Effect of 1-to-1 matching in promoting diversity.

Variant	R_m	mAP_m	# active	% match P	% match GT
-SA+QD (ours)	17.52	23.93	6.43 ± 2.16	0.11 ± 0.13	0.42 ± 0.37
+ 1-to-k matcher [17]	10.38	10.39	20.00 ± 0.00	0.14 ± 0.35	0.06 ± 0.17
+group_matching [7]	17.30	23.71	12.70 ± 6.13	0.15 ± 0.17	0.43 ± 0.36
+hybrid [13]	**17.91**	**24.38**	10.10 ± 6.23	0.14 ± 0.16	0.50 ± 0.36

Table 7. Impact of the proposed architectural modifications.

-SA	+QD	R_m	mAP_m	# active
		16.04	20.84	3.64±1.18
✓		15.31	21.02	3.72±1.16
	✓	16.50	21.43	3.77±1.28
✓	✓	**17.52**	**23.93**	6.43±2.16

Table 8. Effect of the QD dropout rate.

k	R_m	mAP_m
0.00	15.31	21.02
0.25	**17.52**	**23.93**
0.50	0.99	3.84

ing them from collapsing into a redundant prediction. As shown in Tab. 6, replacing it from our (-SA+QD) for a pure 1-to-k matcher [10] leads to redundant predictions, as numerous queries receive the same supervision signal. In contrast, partial relaxations that retain the 1-to-1 matching—e.g., [7, 13]—preserve competition and yield comparable results. This highlights the crucial role of the 1-to-1 matching to ensure that queries that are additionally activated by (-SA+QD) remain diverse and contribute to generalization.

Effect of each component: Table 7 evaluates variants that only include query-dropout (+QD) or only remove self-attention (-SA). Observe that neither component alone yields a significant performance gain as, by themselves, they do not overcome the query collapse—increasing only marginally the number of active queries. Combining both, in turn, increases mAP_m by up to 3.09 while nearly doubling the number of active queries. This confirms the need of solving both coordination and index collapse jointly.

Ablation on the query dropout rate: In Tab. 8 we ablate over the effect of various query-dropout rates—i.e., $k = \{0.0, 0.25, 0.5\}$. Observe that performance peaks at 0.25, after which performance decays. This confirms that a light stochastic regularization encourages broader query utilization, without compromising model convergence.

Scaling the number of potential queries: Figure 7 investigates how the increase of the total number of decoder queries influences the number of active queries as well as performance. The base model quickly saturates. The number of active queries remains nearly constant, and performance severely degrades after peaking at 20 queries. In contrast, our method presents a more steady increase in the number of active queries and performance (mAP_m@0.1). This trend holds up until 20 queries, after which performance stabilizes. Find the extended ablation in Sec. I.

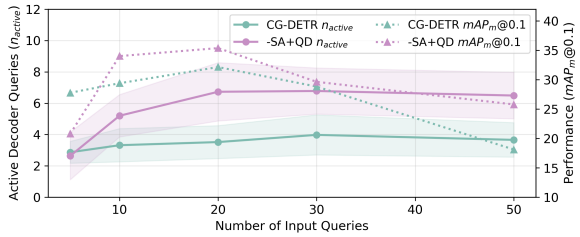

Figure 7. Evolution of the number of active queries and performance over the total number of decoder queries.

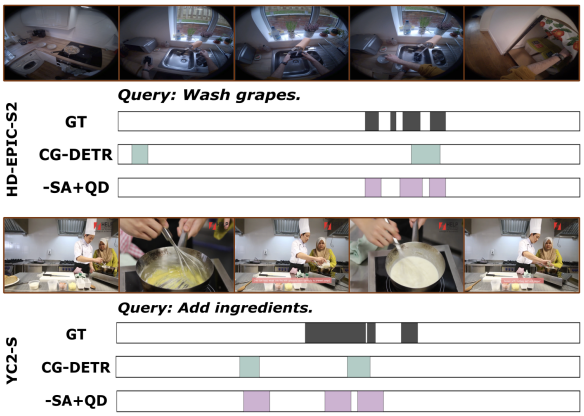

Figure 8. Qualitative results for CG-DETR on HD-EPIC-S2 and YC2-S benchmarks.

Qualitative results: Figure 8 presents several qualitative examples. In these, the base model—CG-DETR (CG)—shows limited success in detecting all GT moments, partially due to its small number of active queries. For example, in the second example, only two queries are activated to retrieve 3 moments. In contrast, our model activates a larger number of queries, leading to a better coverage of the GT moments. See Sec. B for more qualitative examples.

6. Conclusions

In this work, we revisited Video Moment Retrieval (VMR) from the perspective of generalization to search queries, moving beyond existing caption-based datasets. To this end, we introduced three search-query benchmarks, revealing a consistent performance drop when models trained on captions are evaluated on under-specified search queries. We identified and quantified two key factors for this drop: (i) a language gap between detailed and under-specified textual queries, and (ii) a multi-moment gap, arising from the shift from single-moment caption-based queries to multi-moment search queries. We further showed that the latter triggers an active decoder-query collapse. Finally, we introduced various architectural modifications that mitigated this issue, hence improving generalization to search queries—bringing VMR models closer to real-world scenarios.

Acknowledgements

This work has been partially supported by the Spanish project PID2022-136436NB-I00 and by ICREA under the ICREA Academia programme. Research at Bristol is supported by EPSRC Fellowship UMPIRE (EP/T004991/1).

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*, pages 5803–5812, 2017. 2

[2] Peijun Bao, Chenqi Kong, Zihao Shao, Boon Poh Ng, Meng Hwa Er, and Alex C Kot. Vid-morp: Video moment retrieval pretraining from unlabeled videos in the wild. *arXiv preprint arXiv:2412.00811*, 2024. 2

[3] Shyamal Buch, Victor Escorcia, Chuanqi Shen, Bernard Ghanem, and Juan Carlos Niebles. Sst: Single-stream temporal action proposals. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2911–2920, 2017. 2

[4] Zhuo Cao, Heming Du, Bingqing Zhang, Xin Yu, Xue Li, and Sen Wang. When one moment isn't enough: Multi-moment retrieval with cross-moment interactions. *arXiv preprint arXiv:2510.17218*, 2025. 2

[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. 1, 2, 7

[6] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. Temporally grounding natural sentence in video. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 162–171, 2018. 2

[7] Qiang Chen, Xiaokang Chen, Jian Wang, Shan Zhang, Kun Yao, Haocheng Feng, Junyu Han, Errui Ding, Gang Zeng, and Jingdong Wang. Group detr: Fast detr training with group-wise one-to-many assignment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6633–6642, 2023. 2, 8

[8] Xiaoxue Cheng, Junyi Li, Wayne Xin Zhao, Hongzhi Zhang, Fuzheng Zhang, Di Zhang, Kun Gai, and Ji-Rong Wen. Small agent can also rock! empowering small language models as hallucination detector. *arXiv preprint arXiv:2406.11277*, 2024. 3

[9] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 4

[10] Rongyao Fang, Peng Gao, Aojun Zhou, Yingjie Cai, Si Liu, Jifeng Dai, and Hongsheng Li. Feataug-detr: Enriching one-to-many matching for detrs with feature augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(9):6402–6415, 2024. 2, 8

[11] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017. 2, 3

[12] Zhengdong Hu, Yifan Sun, Jingdong Wang, and Yi Yang. Dac-detr: Divide the attention layers and conquer. *Advances in Neural Information Processing Systems*, 36:75189–75200, 2023. 6

[13] Ding Jia, Yuhui Yuan, Haodi He, Xiaopei Wu, Haojun Yu, Weihong Lin, Lei Sun, Chao Zhang, and Han Hu. Detrs with hybrid matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19702–19712, 2023. 2, 8

[14] Jihwan Kim, Miso Lee, and Jae-Pil Heo. Self-feedback detr for temporal action detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10286–10296, 2023. 1, 2

[15] 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*, pages 706–715, 2017. 1, 3, 4, 5

[16] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34: 11846–11858, 2021. 1, 2, 3

[17] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13619–13627, 2022. 2, 8

[18] Pandeng Li, Chen-Wei Xie, Hongtao Xie, Liming Zhao, Lei Zhang, Yun Zheng, Deli Zhao, and Yongdong Zhang. Momentdiff: Generative video moment retrieval from random to real. *Advances in neural information processing systems*, 36:65948–65966, 2023. 2

[19] Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. Towards debiasing sentence representations. *arXiv preprint arXiv:2007.08100*, 2020. 2

[20] Renjie Liang, Chongzhi Zhang, Li Li, Jing Wang, Xizhou Zhu, and Aixin Sun. Tvr-ranking: A dataset for ranked video moment retrieval with imprecise queries. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 231–239, 2025. 2

[21] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 2

[22] Per Linell. *The written language bias in linguistics: Its nature, origins and transformations*. Routledge, 2004. 2

[23] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3651–3660, 2021. 2

[24] WonJun Moon, Sangeek Hyun, Su Been Lee, and Jae-Pil Heo. Correlation-guided query-dependency calibration in video representation learning for temporal grounding. *CoRR*, 2023. 1, 2, 5

[25] WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23023–23033, 2023. 2

[26] Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkilä. Uncovering hidden challenges in query-based video moment retrieval. *arXiv preprint arXiv:2009.00325*, 2020. 2

[27] Yunus Emre Özköse and Pınar Duygulu. Automatic data augmentation for cooking videos. In *2024 32nd Signal Processing and Communications Applications Conference (SIU)*, pages 1–4. IEEE, 2024. 1, 3, 4

[28] Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Kumar Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, et al. Hd-epic: A highly-detailed egocentric video dataset. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23901–23913, 2025. 1, 3, 4

[29] Sandro Pezzelle. Dealing with semantic underspecification in multimodal nlp. *arXiv preprint arXiv:2306.05240*, 2023. 2

[30] Nils Reimers, I Sentence-BERT Gurevych, et al. Sentence embeddings using siamese bert-networks. arxiv 2019. *arXiv preprint arXiv:1908.10084*, 10, 1908. 3

[31] Hassan Sajjad, Patrick Pantel, and Michael Gamon. Underspecified query refinement via natural language question generation. In *Proceedings of COLING 2012*, pages 2341–2356, 2012. 1

[32] Kate Sanders, Reno Kriz, David Etter, Hannah Recknor, Alexander Martin, Cameron Carpenter, Jingyang Lin, and Benjamin Van Durme. Grounding partially-defined events in multimodal data. *arXiv preprint arXiv:2410.05267*, 2024. 2

[33] Hao Sun, Mingyao Zhou, Wenjing Chen, and Wei Xie. Tr-detr: Task-reciprocal transformer for joint moment retrieval and highlight detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4998–5007, 2024. 1, 2

[34] Jiajin Tang, Zhengxuan Wei, Yuchen Zhu, Cheng Shi, Guanbin Li, Liang Lin, and Sibei Yang. Sim-detr: Unlock detr for temporal sentence grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22760–22771, 2025. 2

[35] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024. 3

[36] Y Wang, X Zhang, T Yang, and J Sun. Anchor detr: Query design for transformer-based object detection. arxiv 2021. *arXiv preprint arXiv:2109.07107*, 3, 2021. 2

[37] Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. Multilevel language and vision integration for text-to-clip retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9062–9069, 2019. 2

[38] Lizhen Xu, Shanmin Pang, Wenzhao Qiu, Zehao Wu, Xiuxiu Bai, Kuizhi Mei, and Jianru Xue. Redundant queries in detr-based 3d detection methods: Unnecessary and prunable. *arXiv preprint arXiv:2412.02054*, 2024. 2

[39] Yifang Xu, Yunzhuo Sun, Benxiang Zhai, Ming Li, Wenxin Liang, Yang Li, and Sidan Du. Zero-shot video moment retrieval via off-the-shelf multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8978–8986, 2025. 2

[40] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. Deconfounded video moment retrieval with causal intervention. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 1–10, 2021. 2

[41] Sunjae Yoon, Ji Woo Hong, Eunseop Yoon, Dahyun Kim, Junyeong Kim, Hee Suk Yoon, and Chang D Yoo. Selective query-guided debiasing for video corpus moment retrieval. In *European Conference on Computer Vision*, pages 185–200. Springer, 2022. 5

[42] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, pages 492–510. Springer, 2022. 2

[43] Chuyang Zhao, Yifan Sun, Wenhao Wang, Qiang Chen, Errui Ding, Yi Yang, and Jingdong Wang. Ms-detr: Efficient detr training with mixed supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17027–17036, 2024. 2, 8

[44] Pengcheng Zhao, Zhixian He, Fuwei Zhang, Shujin Lin, and Fan Zhou. Ld-detr: Loop decoder detection transformer for video moment retrieval and highlight detection. *arXiv preprint arXiv:2501.10787*, 2025. 1, 2, 5

[45] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 3

[46] Benjin Zhu, Zhe Wang, Shaoshuai Shi, Hang Xu, Lanqing Hong, and Hongsheng Li. Conquer: Query contrast voxel-detr for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9296–9305, 2023. 2

[47] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2