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

Supplementary Material

This supplementary material complements the main text showing additional details and ablation studies. Concretely, Sec. A expands on the provided description of the two introduced metrics, as well as the underlying intuition on situations where standard VMR metrics fail to provide a complete evaluation. Sec. B presents various additional qualitative results for each of the proposed benchmarks, and Sec. C further implementation details regarding various aspects like the search-query pipeline, the evaluation setup, or details regarding the models and their optimization. Sec. D expands on the main results presented in the main text, including the study of an oracle model trained on search queries and for comparability purposes, the performance of these models in terms of the standard mAP or the degradation study of non-DETR architectures. Sec. E provides a comprehensive study of the realism and similarity of our proposed search queries with respect to real life queries. Sec. F studies the relationship between calibration and the observed active decoder query collapse. Sec. G expands on the results presented in the main text, disentangling where the performance gains of (-SA+QD) come from by evaluating the models on single-moment, and multi-moment queries, independently. Sec. H additionally provides further details on the impact of both language and multi-moment gap, while Sec. I similarly expands the results of all the ablation studies presented in the main text. Finally, Sec. J conducts a qualitative study of the generated search queries for each of the proposed benchmarks.

A. Expanded description metrics

In this section we provide further details of the two metrics that we introduce in this work (see Sec. 3 of the main text), namely R_m and mAP_m. Concretely, we provide additional intuition including various illustrative examples to better understand the pitfalls of existing metrics, and hence, the need of our proposed metrics. We moreover describe in further detail the formalization of both metrics.

A.1. Intuition

Observe in Fig. A an illustration of the behavior of our proposed metric R_m with respect to the more standard $R1$. In the left example, observe that the highest confidence prediction matches one of the GT moments, while the second GT is matched by the third highest-prediction. In this case, standard R1 metric would predict a score of 1, since the top prediction does correspond to one of the GT, but this score does not account for the quality of the detection of the second GT moment. R_m in contrast, would provide a per-GT score, where the first moment would get a score of 1 since it was matched to the highest-confidence prediction, while the second GT moment—matched to the third highest confidence—would get a score of 0. The reason is that the model ranked the second prediction with a higher confidence, which corresponds to a false-positive—not matching any GT. Following the intuition of R1, this prediction has not been "accurately" retrieved.

The contrary happens in the second example, where R_m assigns a score of 1 for both GT, since one corresponds to the highest confidence, and the second, despite being ranked second, it is not penalized since the prediction with a higher confidence is also a match to a different GT. This hence shows similar behavior to R1.

Finally, in the third example both $R1$ and R_m have a similar behavior. Since a false-positive prediction is ranked on top, all the remaining predictions that correspond to a GT get assigned a score of 0. These examples exemplify how R_m computes per-GT scores, evaluating the quality of the retrieval of each GT moment independently, and without being affected by other potential matches that may co-occur.

A similar behavior is shown in Fig. B, which showcases the behavior of mAP_m with respect to mAP. This considers a scenario where the model detects a given GT moment with the highest-confidence prediction, while the second is detected with the lowest one. In this case, mAP computes a global score of 0.7, where the correct prediction of one of the GT moment masks the poor detection of the other. mAP_m, in contrast, computes a score for each GT were the

*Work partially completed whilst at University of Bristol.

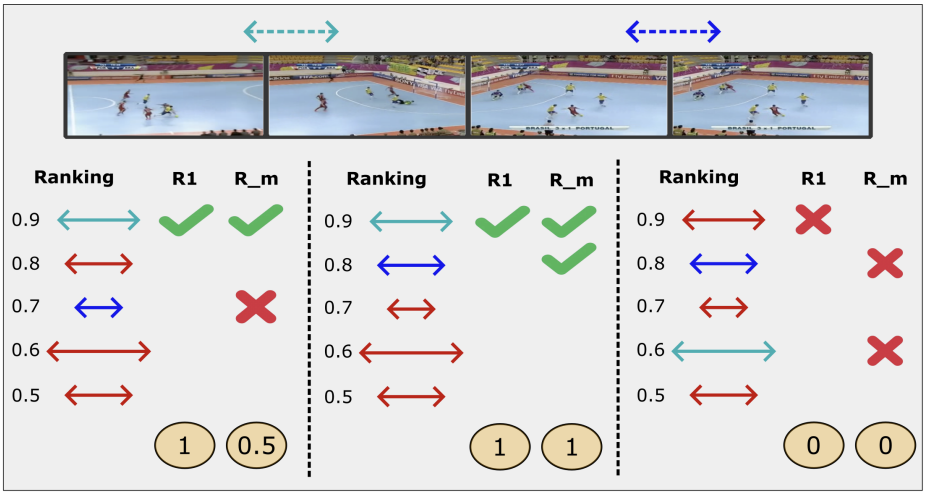

Figure A. Intuitive examples that showcase the behavior of both R_m and $R1$. Here the solid lines correspond to prediction, and the dashed ones correspond to GT moments. Moreover, in R1, the checkmark indicates that the entire query is marked as correct, while for R_m, since it performs a per-GT evaluation, this indicates whether the corresponding prediction "correctly" retrieves a GT moment or not, based on the criterion defines by the metric. The orange circles indicate the global score of the instance, which is consistent with the single score produced by $R1$, or by the average of the multiple per-GT scores for R_m.

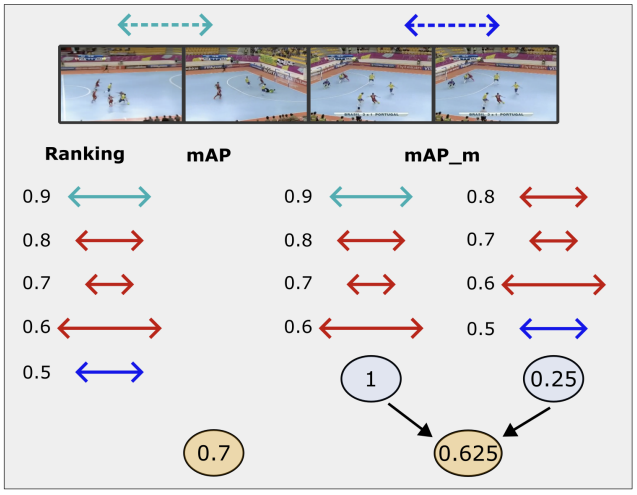

Figure B. Example that showcases the behavior of both mAP and mAP_m. mAP computes a global score for the entire ranking, which is depicted in its corresponding orange circles. mAP_m, in turn, computes a score for each of the two GT moments separately. The example shows the two rankings that this evaluation leverages, effectively ignoring the influence of predictions that, while match a different GT, should not be considered invalid. The orange circle from mAP_m corresponds to the average of the two respective per-GT scores. .

only difference is that each evaluation ignores all the predictions that match any other GT. For instance, to evaluate the dark blue moment (right most example), mAP_m ignores the prediction that corresponds to the light blue one. This avoid penalizing matches that, while different, are still valid and should thus not be considered incorrect. Hence, in this case, the score for one of the moments is 1, while the other is of 0.25 giving a final mAP_m score of 0.625 when averaged across the two GT moments.

A.2. Preliminaries

Given a video-query pair, a VMR model outputs a set of K predictions—i.e., candidate moments—, denoted as:

$$\mathcal{P} = \{p_1, \dots, p_K\}, \tag{1}$$

where each prediction p_i is a temporal segment predicted by the model, associated with a confidence score $c(p_i)$. These predictions are sorted in descending order:

$$c(p_1) \geq c(p_2) \geq \cdots \geq c(p_K). \tag{2}$$

Moreover, this video-query pair maps to a set of GT moments $\mathcal{G}$:

$$\mathcal{G} = \{g_1, \dots, g_n\} \tag{3}$$

Given a certain IOU threshold τ, we follow the existing literature [4] and define a match of a prediction with a GT moment as:

$$match(p_i, g_j, \tau) = \begin{cases} 1 & IOU(p_i, g_j) \geq \tau \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

For convenience, let us also define the cases where a prediction matches a GT moment that while valid, differs to the moment g_j that is under evaluation:

$$match\_other(p_i, g_j, \tau) = \begin{cases} 1 & \exists g_k \neq g_j \ st. \ IOU(p_i, g_k) \geq \tau \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

A.3. Multi-moment recall R_m

Standard Recall@1 (R1) assigns a score per video-query pair, this being 1 if the highest-ranked prediction matches any of the GT moments, and 0 otherwise. This provides only a partial performance overview when evaluating multi-moment queries—those matching to multiple GT moments—as this metric does provide information on whether the model was able to successfully detect all the GT moments.

The goal of the R_m metric is to instead evaluate the detection quality for each of the GT moments, independently. Importantly, our metric avoids the interference of other co-occurring moments in the score assigned to the evaluation of a given GT moment.

More specifically, R_m considers a given GT moment g_j as correctly retrieved if it appears before any false positive predictions, ignoring predictions that do not match g_j as they cannot be considered mistakes, since they match different, equally valid, GT moments.

Formally, let us define the index of the first prediction matching g_j:

$$i^* = \min(i \mid match(p_i, g_j, \tau) = 1). \tag{6}$$

Moreover, the index of the first false-positive is defined as:

$$i_j^{FP} = min(i|match(p_i, g_j, \tau) = 0 \wedge match\_other(p_i, g_j, \tau) = 1). \tag{7}$$

With this, the recall score for a given GT moment g_j is defined as follows:

$$R_m(g_j, \tau) = \begin{cases} 1 & i^* \geq i_j^{FP} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

Finally, in order to obtain global scores for our dataset, we additionally compute an aggregated score:

$$R_m(\tau) = \frac{1}{|\mathcal{G}|} \sum_{g_j \in \mathcal{G}} R_m(g_j, \tau). \tag{9}$$

Note that similarly to other metrics, we still obtain a dataset-level metric, however, this score assigns an equal weight to all the GT moments in the dataset, regardless of the number of moments that co-occur with it in the same video-query pair. This is key as otherwise, a multi-moment query comprising of 10 GT moments would have the same weight as that of a single query mapping to a single moment. Fixing this issue is key to ensure a fair comparison across levels of specificity, as well as in general, to provide a more fine-grained evaluation that looks at performance from a per-GT perspective, instead of a query-level one.

A.4. Multi-moment mAP (mAP_m)

Similarly to R1, mAP has the fundamental limitation that it also produces a query-video level score. Hence, it obscures the performance of the potentially multiple GT moments that may correspond to such query, as the good detection of a given moment can mask poor detections or even GT moments that were not detected at all. As argued in Sec. 3, this breaks comparability in our setup, even though we argue that this limitation also extends to the evaluation of multi-moment queries in general.

To overcome this issue, similarly to R_m we propose evaluating the detection performance of each of the GT moments, independently, ensuring that a good/bad detection on one GT moment does not interfere with the scores of any other co-occurring moments.

Accordingly, for a given GT moment g_j, we define the set of true positives TP predictions as:

$$TP_j = \{p_i \mid match(p_i, g_j, \tau) = 1\}. \tag{10}$$

The false positives, being the predictions that do not match any GT moment is defined as:

$$FP_j = \{p_i \mid match(p_i, g_j, \tau)) = 0 \wedge match\_other(p_i, g_j, \tau) = 0\}. \tag{11}$$

And finally, we define the set of predictions that are ignored since, even though they do not match the moment that is currently evaluated (g_j), they nevertheless match another valid moment (not g_j). Ignoring them prevents these predictions from penalizing the metric for g_j, as this should neither benefit this metric, nor penalize it as a mistake, when it is a perfectly valid prediction. Formally,

$$IGN_j = \{p_i \mid match\_other(p_i, g_j, \tau) = 1\} \tag{12}$$

With this, for each of the GT moments g_j we compute the corresponding precision P_j and recall R_j:

$$P_j(k) = \frac{\text{\#TP up to rank } k}{\text{\#TP up to rank } k + \text{\#FP up to rank } k} \tag{13}$$

$$R_j(k) = \frac{\text{\#TP up to rank } k}{1}, \tag{14}$$

and following the original work [4], compute the corresponding area-under-the-curve (AUC):

$$AP_m(g_j, \tau) = \text{AUC}(P_j, R_j) \tag{15}$$

Similarly to R_m, we also compute a dataset level score as:

$$mAP_m(\tau) = \frac{1}{|\mathcal{G}|} \sum_{g_j \in \mathcal{G}} AP_m(g_j). \tag{16}$$

This again results in a score where each of the GT moments in $\mathcal{G}$ have an equal weight, making comparisons across levels of specificity fair, as the score of the detection quality for a given GT moment g_j is independent to the moments that it co-occurs with.

B. Qualitative results VMR

Find below various qualitative examples that showcase the performance of our proposed modification (-SA+QD) with respect to its corresponding baseline, CG-DETR.

Concretely, Fig. C and Fig. D show two different examples for each of the scenarios included in our proposed benchmarks—i.e., HD-EPIC-S1/S2/S3, YC2-S and ANC-S. Observe that in numerous examples, the base CG-DETR is unable to activate sufficient predictions with a non-vanishing confidence, which hinders the capacity to detect multi-moment queries as the number of active queries is smaller than the number of GT moments to retrieve. This is considerably mitigated by (-SA+QD) that consistently activates more decoder queries, thus showing considerably better behavior in these scenarios.

C. Implementation details

In this section we provide additional details required for reproducibility. These include the search-query pipeline, the three proposed search-query benchmarks, and the implementation and optimization details of the evaluated baselines.

C.1. Search-query pipeline

C.1.1. Under-specification stage

Rewriter: As described in the main text, we obtain the under-specifications of the original caption-based queries using an LLM agent based on Gemma3-12b-it. We choose this model due to its open-source availability and its good performance for this task compared to other LLMs we evaluated.

The rewriter follows an in-context learning strategy as a way to guide the agent towards the desired levels of specificity. Concretely, for each of the benchmarks we pass a small set of examples that are manually annotated by a human annotator. It is key that these examples are benchmark-specific so as to avoid the shift between the distribution of the in-context examples and that of the instances that we aim to under-specify. In Fig. E, Fig. F, Fig. G, Fig. H and Fig. I we include the prompts used for each of the benchmarks:

Validator: To ensure the quality of the rewritings, each under-specified query together with its corresponding caption-based query is passed through a validator. This validator, based on an identical Gemma3-12b-it LLM agent, determines if the under-specified query is consistent with the caption. Valid instances are subject to a random validation process by human annotators, while the invalid instances are all manually reviewed and rewritten by a human annotator, if necessary. Note that thanks to the curated in-context examples, the number of invalid instances are few dozens per benchmark, representing less than 1% of the queries. Find in Fig. J the prompt used for the validator.

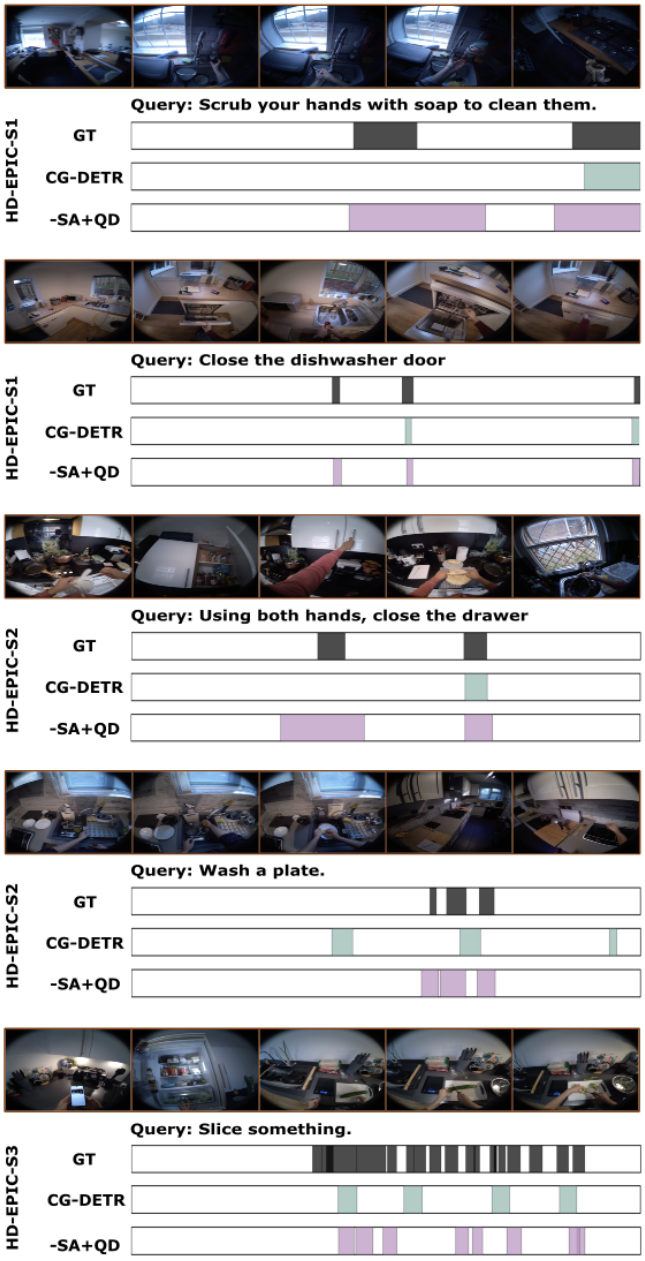

Figure C. Qualitative results of the performance of CG-DETR and our proposed modification (-SA+QD) on HD-EPIC-S1/S2/S3.

C.2. Grouping stage

The grouping stage is further divided into two different steps:

Similarity-based grouping: For each video, we compute the STSB-Roberta-Large sentence embeddings [12] of all the queries that occur in this video. We then compute their pairwise cosine similarities and form a graph where the nodes are the queries, and where two nodes are connected if their corresponding queries present a similarity equal or

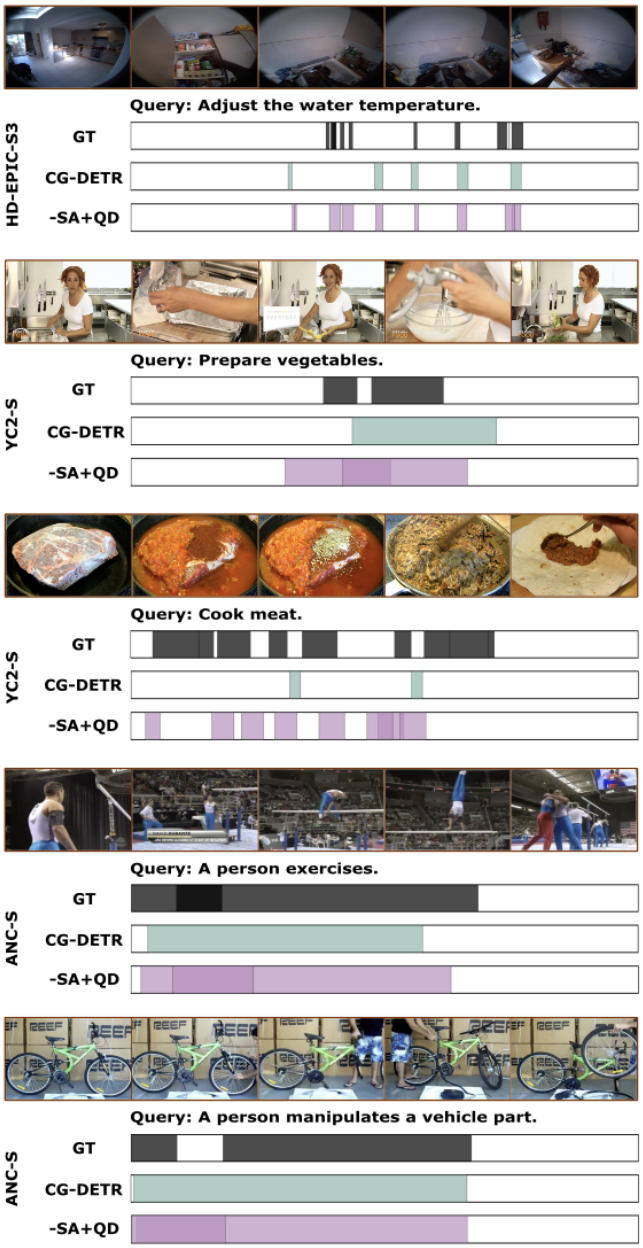

Figure D. Qualitative results of the performance of CG-DETR and our proposed modification (-SA+QD) on HD-EPIC-S3, YC2-S and ANC-S.

```
Simplify this query by removing or generalizing
    unnecessary information, but keeping the
    meaning of the sentence.
Format your output as:
Simplified query: <your simplified query>

Example 1:
Query: Throw the big orange into the green
    recyclable bag using the left hand. With the
    right hand, pick up the end of the green roll
     and pull the roll outwards to pull one bag
    out of the roll using the right hand.
Simplified query: Throw the orange into the bag.
    Then pick up the roll and use it to pull a
    bag out.

Example 2:
Query: Open the upper cupboard by holding the
    handle of the cupboard with the left hand.
Simplified query: Open the cupboard.

Example 3:
Query: Switch the button of the socket using the
    left hand. This enables the power to access
    the food processor so as to start it.
Simplified query: Switch the button to start the
    food processor.

Example 4:
Query: Turn the dial of the food processor by
    turning it clockwise to switch it on. The
    juicer will now start rotating
Simplied query: Turn the dial by turning it to
    switch it on.

Example 5:
Query: Using the left hand, remove the plastic
    cover of the blue scissors. This action
    occurs at the periphery, quite off-screen.
Simplified query: Remove the cover of the
    scissors.


Now, simplify this query:
Query: {query}
Simplified query:
```

Figure E. Rewriter prompt for HD-EPIC-S1.

greater to 0.85. Given these connections, we use a DFS algorithm to form the groups/clusters that essentially contain the same semantics.

Representative search query generation: Since members of the same group/cluster can still present some minor differences in their corresponding under-specified queries, we generate a single representative search query that corresponds to all of them. For instance, consider two very similar under-specifications *"hold the pan"* and *"hold the pot"*. Despite their similarity, the first includes unique information, that is not present in the seconds, and vice-versa. Thus, to create a single search query that corresponds to all the members of the group—i.e., containing shared semantics only—we employ a final identical Gemma3-12b-it agent that takes all the corresponding under-specifications of a given group and computes their final search query—e.g., *"hold the kitchenware"* following the previous example. See in Fig. K the prompt that we used.

```
Simplify this query by removing or generalizing
    unnecessary information, but keeping the
    meaning of the sentence. Avoid using very
    vague words like ``something'' or ``items''.
Format your output as:
Simplified query: <your simplified query>

Example 1:
Query: Throw the big orange into the green
    recyclable bag using the left hand. With the
    right hand, pick up the end of the green roll
     and pull the roll outwards to pull one bag
    out of the roll using the right hand.
Simplified query: Throw an orange away.

Example 2:
Query: Open the upper cupboard by holding the
    handle of the cupboard with the left hand.
Simplified query: Open a cupboard.

Example 3:
Query: Switch the button of the socket using the
    left hand. This enables the power to access
    the food processor so as to start it.
Simplified query: Switch a button on.

Example 4:
Query: Turn the dial of the food processor by
    turning it clockwise to switch it on. The
    juicer will now start rotating
Simplied query: Turn a dial on.

Example 5:
Query: Using the left hand, remove the plastic
    cover of the blue scissors. This action
    occurs at the periphery, quite off-screen.
Simplified query: Remove a cover.


Now, simplify this query:
Query: {query}
Simplified query:
```

Figure F. Rewriter prompt for HD-EPIC-S2.

```
Simplify this query by removing or generalizing
    unnecessary information, but keeping the
    meaning of the sentence.
Format your output as:
Simplified query: <your simplified query>

Example 1:
Query: Throw the big orange into the green
    recyclable bag using the left hand. With the
    right hand, pick up the end of the green roll
     and pull the roll outwards to pull one bag
    out of the roll using the right hand.
Simplified query: Throw items away.

Example 2:
Query: Open the upper cupboard by holding the
    handle of the cupboard with the left hand.
Simplified query: Open an item.

Example 3:
Query: Switch the button of the socket using the
    left hand. This enables the power to access
    the food processor so as to start it.
Simplified query: Switch something on.

Example 4:
Query: Turn the dial of the food processor by
    turning it clockwise to switch it on. The
    juicer will now start rotating
Simplied query: Turn something on.

Example 5:
Query: Using the left hand, remove the plastic
    cover of the blue scissors. This action
    occurs at the periphery, quite off-screen.
Simplified query: Remove an item.


Now, simplify this query:
Query: {query}
Simplified query:
```

Figure G. Rewriter prompt for HD-EPIC-S3.

C.3. Evaluation setup

Below we provide additional details on the evaluation of each of the benchmarks:

HD-EPIC-S1/S2/S3: We extract InternVideo2 [13] features at an FPS rate of 3. Because HD-EPIC contains very long videos, we find that a naive evaluation on the entire videos yields uninformative results given the extremely low baseline scores. Since long-video VMR detection is beyond the scope of this paper, even though this constitutes a promising line of research, we trim the videos into 500-frame windows, treating each of them as independent instances. We further discard the windows that do not contain any GT moment, avoiding the important issue of dealing with the assumption that every instance contains at least one GT moment [5]. This is a relevant issue that falls beyond the scope of this paper, mainly because the majority of baselines, including the ones we evaluate on do not provide built-in mechanisms to deal with these situations.

Moreover, HD-EPIC does not provide pre-defined data splits suitable for VMR. Hence, we construct a training and testing split using an 80-20 split of the per-participant original data.

YC2-S: We extract InternVideo2 features at an FPS rate of 3, and use the original training and validation splits as proposed by [16].

ANC-S: We extract InternVideo2 features at an FPS rate of 3, and use the original training and validation splits as proposed by [8].

```
Simplify this query by removing or generalizing
    unnecessary information, but keeping the core
     meaning of the sentence.
Format your output as:
Simplified query: <your simplified query>

Example 1:
Query: pick the ends off the verdalago
Simplified query: Add ingredient.

Example 2:
Query: pour the dressing over the salad and mix
Simplified query: Make a salat

Example 3:
Query: chop lettuce and place it in a bowl.
Simplified query: Prepare vegetables.

Now, simplify this query:
Query: {query}
Simplified query:
```

Figure H. Rewriter prompt for YC2-S.

```
Simplify this query by removing or generalizing
    unnecessary information, so that the
    simplified query can match other overspecific
     ones if possible.
Format your output as:
Simplified query: <your simplified query>

Example 1:
Query: He then bends down and grabs a ball.
Simplified query: A person manipulates an object.

Example 2:
Query: Then one man stands in a field holding a
    wooden object and begins twisting it.
Simplified query: A person manipulates an object.

Example 3:
Query: There was a penalty and one players
    attempts to hit the ball into the goal from
    the side.
Simplified query: People play a game.

Example 4:
Query: A group of people holding paintball guns
    and dressed in costume run into a staged
    setting as if in combat.
Simplified query: People play a game.

Now, simplify this query:
Query: {query}
Simplified query:
```

Figure I. Rewriter prompt for ANC-S.

C.4. Models and optimization

As described in the main text, we select CG-DETR and LD-DETR as representatives of the DETR-based VMR models. These models are trained on a single NVIDIA GeForce RTX 3090. We use all the original hyper-parameters from [11] and [15], respectively, across all the benchmarks. Our method (-SA+QD) only introduces one new hyperparameter, being the QD rate, selected via grid search and kept as 0.25 across all the benchmarks and models.

```
Determine whether this simplified query is a
    valid underspecification of the original
    query or it is a hallucination that describes
     something different.
Format your output as:
Validity: <valid | invalid>

Example 1:
Original: Throw the big orange into the green
    recyclable bag using the left hand.
Simplified: Throw the orange into the bag.
Validity: valid

Example 2:
Original: Switch the button of the socket using
    the left hand.
Simplified: A dog jumping a fence.
Validity: invalid

Now, evaluate this pair:
Original: {original}
Simplified: {simplified}

Validity:
```

Figure J. Instruction of the Validator.

```
Combine these queries into a single unified
    description that applies to all the queries.
    Avoid adding unncessary ands/ors, trying to
    make it as concise as possible:
Example 1:
Queries: [Hold the pan using the left hand to
    cook tomatoes, Hold the pot using the right
    hand to cook onions]

Unified description: Hold the kitchenware using a
     hand to cook vegatables.
Example 2:
Queries:  [Take off the left glove. , Take off
    the right glove]
Unified description: Take off the glove.
Now create a unified description for the
    following queries:
Queries: {ann['original_queries']}
Unified description:
```

Figure K. Prompt of the agent computing the representative search query for each of the clusters.

D. Experimental results

D.1. Expanded main results

In this section we expand upon the main results presented in the main text. Although, as argued in the main text, the standard mAP metric does not provide a fair evaluation of our setup, we find it helpful to include these results completeness and comparability with prior work. As shown in Tab. A, Tab. B and Tab. C, our method also achieves consistent gains in mAP across all the benchmarks.

Table A also reports the results of an oracle model. The oracle corresponds to training the base architectures—CG-DETR and LD-DETR, respectively— directly on the data at the target specificity. For example, for HD-EPIC-S2, the oracle is obtained by training the baseline model on the training split of HD-EPIC-S2 derived from our proposed search-query pipeline. While this is not aligned with the underlying goal of this work, this being training on the standard captions while generalizing to more under-specified search queries, this still provides a meaningful upper-bound on the achievable performance.

Interestingly, our proposed (+SA-QD) significantly closes the gap between base and oracle model. For instance, on HD-EPIC-S2, our proposal closes the gap by up to 82% of mAP_m@0.1.

D.2. Generalization across different architectures

To verify that the performance gap is not exclusive to the DETR family, we extend our evaluation to Flash-VTG [1], a state-of-the-art anchor-based architecture. Unlike previous models that localize via a Transformer decoder and learnable queries, Flash-VTG utilizes a *Temporal Feature Layering* (TFL) module and *Adaptive Score Refinement* (ASR) to regress moments from multi-scale anchors. This fundamental architectural shift allows us to test if the observed degradation is a DETR-specific artifact or a systemic challenge of the VMR task.

Performance on under-specified queries: We evaluated Flash-VTG on our proposed benchmarks (Table D). Despite its use of fixed anchors rather than learnable queries, Flash-VTG experiences a similar performance collapse when moving from descriptive captions to realistic search queries. On *HD-EPIC-S1*, for instance, we observe a drop of 12.22% in mAP@0.1 (from 43.70 to 31.48) and 9.11% in R@0.1 (from 30.56 to 21.45). This trend is even more pronounced on *YC2-S*, where mAP@0.5 plummets by 34.46%. These results confirm that the "visual bias" inherent in current datasets affects models regardless of whether they are query-based or anchor-based.

By demonstrating that the performance gap persists in a completely different architectural paradigm, we provide strong evidence that the bottleneck lies in the training data distribution rather than specific architectural choices. We leave as future work studying potential mitigation strategies for this family of methods.

E. Study of the realism of the queries

E.1. Alignment with user queries

To ensure our generated benchmarks adequately simulate real-world search queries, we explore two different analysis:

Linguistic comparison: We perform a linguistic comparison against MS-MARCO [3], a dataset of real user search logs. Our analysis shows that standard caption-based queries in HD-EPIC are significantly more descriptive than real search queries, being approximately 3x longer (17.7 vs. 5.9 words) and over-saturated with adjectives (8.1%). Importantly, HD-EPIC-S1 queries successfully bridge this gap through our proposed under-specification pipeline. This allows, for instance, matching the MS-MARCO average length (5.98 words vs. 5.89 words) and verb density (22.6% vs. 21%).

User realism study: Additionally to the aforementioned quantitative analysis, we conducted a study with 22 participants who rated 20 samples each for realism. For each of the selected search queries, they were asked: *"Is this search query realistic? In other words, could you imagine yourself typing this into a search bar to find this specific moment (caption provided as context)?"*. Participants gave *HD-EPIC-S1* a realism score of 89%, confirming that the queries effectively simulate actual user behavior.

E.2. Semantic fidelity and grouping reliability

We further validated that the under-specification process (see Sec. 3.2.1) does not lose the core intent of the original captions or introduce noise into the retrieval task through three main metrics:

Intent preservation: Using dependency parsing to extract root verbs and nouns, we verified that 96.9% of the queries in *HD-EPIC-S1* preserve the original intent of the source caption. Here a search query is considered to preserve the core intent of its given caption if it contains both its main verb and noun (or a synonym). This ensures that the simplified search queries remain semantically grounded in the ground-truth video content.

Discriminability: To confirm that under-specification does not induce ambiguity, we measured the cosine similarity of the search queries against their corresponding captions using *all-MiniLM-L6-v2* embeddings. The queries yielded a high similarity of 0.89 against their own captions compared to 0.21 against unrelated ones, providing a clear 0.68 margin that ensures no induced confusion.

Grouping reliability: For multi-moment instances, we validate that the clusters resulting from the query grouping stage present a high intra-cluster similarity (0.95) and a low

Table A. 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	mAP_m @0.1	mAP_m @0.3	mAP_m @0.5	mAP @0.1	mAP @0.3	mAP @0.5
CG-DETR	Original	base	34.44	21.63	11.32	42.69	26.96	14.26	42.69	26.96	14.26
		-SA+QD	34.24	22.68	12.59	45.79	30.52	17.16	45.79	30.52	17.16
	S1	base	28.61	17.95	8.99	36.21	22.84	11.59	38.52	24.33	12.27
		-SA+QD	29.87	19.69	10.86	39.74	26.49	14.87	41.65	27.98	15.85
		base (oracle)	28.85	17.44	9.08	40.42	25.12	13.1	42.58	26.74	14.02
	S2	base	24.71	15.52	7.89	32.15	20.1	10.29	34.19	21.29	10.79
		-SA+QD	26.17	17.00	9.40	35.38	23.39	13.04	36.97	24.68	13.80
		base (oracle)	27.34	17.47	8.90	39.82	25.79	13.22	42.13	27.31	14.00
	S3	base	9.50	4.61	2.08	16.20	8.01	3.58	20.99	11.57	5.29
		-SA+QD	10.57	6.52	3.45	17.27	10.65	5.54	22.07	14.19	7.86
		base (oracle)	12.31	6.09	3.06	23.29	10.94	5.11	30.56	15.3	7.37
LD-DETR	Original	base	34.75	23.90	13.46	42.59	29.17	16.42	42.59	29.17	16.42
		-SA+QD	35.33	24.51	13.37	46.83	32.52	18.01	46.83	32.52	18.01
	S1	base	29.42	19.77	10.50	36.55	24.5	13.18	38.94	26.25	14.29
		-SA+QD	30.18	20.26	10.83	40.50	27.54	14.94	42.58	29.15	16.08
		base (oracle)	29.92	18.61	8.74	41.5	26.33	12.78	43.73	28.00	13.61
	S2	base	25.23	16.38	8.46	32.42	21.11	10.93	34.32	22.58	11.89
		-SA+QD	26.36	16.98	8.87	36.37	23.75	12.54	38.24	25.20	13.56
		base (oracle)	29.78	20.82	11.76	41.93	29.43	16.77	44.11	31.16	17.85
	S3	base	10.44	5.37	2.58	16.48	8.65	4.11	18.63	9.41	4.32
		-SA+QD	10.44	5.28	2.39	17.79	9.06	4.19	21.02	10.11	4.47
		base (oracle)	10.35	5.61	2.67	20.59	11.14	5.16	26.97	13.28	5.63

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

Model	Input	Variant	R_m @0.1	R_m @0.3	R_m @0.5	R_m @0.75	mAP_m @0.1	mAP_m @0.3	mAP_m @0.5	mAP_m @0.75	mAP @0.1	mAP @0.3	mAP @0.5	mAP @0.75
CG-DETR	Orig.	base	63.32	50.71	37.79	20.11	69.47	55.55	41.04	17.18	69.47	55.55	41.04	17.18
		-SA+QD	62.46	51.05	37.19	20.27	70.25	57.97	42.79	18.05	70.25	57.97	42.79	18.05
	S	base	28.92	19.87	11.22	4.60	38.83	26.96	15.21	4.07	47.74	33.61	19.40	5.27
		-SA+QD	29.97	20.32	11.38	4.36	41.00	29.4	17.21	4.65	49.52	36.41	21.86	6.33
LD-DETR	Orig.	base	68.06	56.34	39.75	19.69	73.15	60.62	42.79	15.88	73.15	60.62	42.79	15.88
		-SA+QD	70.20	55.66	37.06	17.63	76.35	61.71	42.01	15.05	76.35	61.71	42.01	15.05
	S	base	33.13	23.48	11.70	4.44	41.69	30.04	15.58	4.09	51.86	37.85	20.0	5.45
		-SA+QD	35.86	24.76	13.17	5.15	45.66	33.09	18.74	4.90	56.05	41.26	23.89	6.14

inter-cluster similarity (0.31). This indicates that only relevant captions are clustered together without "blurring" the boundaries with other action queries.

F. Impact of calibration in the query collapse

In this section we examine if existing confidence calibration methods could resolve the query collapse issue. To this end, Fig. L reports, for each of the learnable queries, the correlation between its regression quality—i.e., measured as the proportion of times it achieves an IOU of at least 0.1 with a GT segment—and its confidence score.

The plot reveals that the issue does not stem from confidence scores that fail to reflect the true quality of the regression estimate—i.e., marking as inactive, queries that in fact produce accurate predictions. Instead, what we observe is that inactive queries—i.e., with low confidence scores—produce substantially worse regression estimates. Thus, to some extent, confidence scores do capture the true quality of the regression estimate. Therefore, the core problem lies not in calibration, but in the lack of mechanisms to encourage more queries to produce accurate moment predictions.

To further support this claim, in Tab. E we evaluate several confidence calibration mechanisms [7, 10]. These results demonstrate how these methods actually lead to a performance degradation, reinforcing that calibration alone cannot overcome active-query collapse.

G. Extended results for disentanglement between "single" and "multi"

In this section we provide further details on the ablation analysis of the main text that aims evaluate separately the performance of single-moment and multi-moment queries. Concretely, in Tab. F we disentangle the performance of single and multi-moment queries, independently, for CG-DETR evaluated on HD-EPIC-S1/S2/S3. Likewise, Tab. G and Tab. H disentangle the performance of CG-DETR on single and multi-moment queries when evaluated on YC2-S and ANC-S benchmarks, respectively.

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

Model	Input	Variant	R_m				mAP_m				mAP			
			@0.1	@0.3	@0.5	@0.75	@0.1	@0.3	@0.5	@0.75	@0.1	@0.3	@0.5	@0.75
CG-DETR	Orig.	base	75.48	60.00	44.02	26.21	82.31	69.36	53.15	25.7	82.31	69.36	53.15	25.7
		-SA+QD	75.19	60.30	44.96	26.47	82.25	69.91	54.5	25.67	82.25	69.91	54.5	25.67
	S	base	60.44	40.89	24.56	12.97	72.18	54.9	36.41	15.07	73.34	55.84	37.49	15.71
		-SA+QD	63.75	43.12	25.50	13.36	74.00	56.42	37.2	15.09	75.54	57.52	38.52	15.78
LD-DETR	Orig.	base	75.72	60.63	45.17	27.15	82.73	70.3	54.46	26.62	82.73	70.3	54.46	26.62
		-SA+QD	76.72	61.44	45.68	27.87	83.15	70.97	55.52	27.29	83.15	70.97	55.52	27.29
	S	base	62.58	43.00	26.08	13.92	73.35	56.17	36.79	15.16	74.65	57.15	37.93	15.93
		-SA+QD	65.21	43.89	25.77	13.36	74.25	56.31	36.69	15.15	76.13	57.58	37.88	15.88

Table D. Results of Flash-VTG on all our proposed benchmarking scenarios.

Benchmark	Input	R_m			mAP_m		
		@0.1	@0.3	@0.5	@0.1	@0.3	@0.5
HD-EPIC	Original	30.56	23.83	16.46	43.70	36.08	26.89
	S1	21.45	16.50	10.98	31.48	25.71	18.83
	S2	16.97	13.47	8.83	25.93	21.48	15.58
	S3	5.64	4.59	3.07	9.63	8.01	5.91
YC2-S	Original	77.66	66.09	49.31	84.78	75.02	60.22
	S	39.86	28.46	17.41	49.78	38.34	25.76
ANC-S	Original	74.95	59.12	43.71	80.93	68.57	55.26
	S	54.05	36.37	22.52	64.63	48.79	34.45

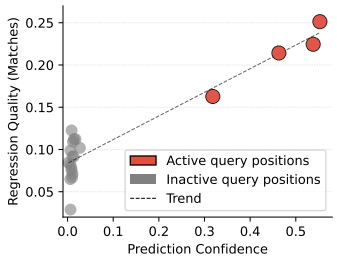

Figure L. Correlation between the average ratio of matched predictions—i.e., predictions with an IOU of at least 0.1 with one of the GT moments—with respect to their respective confidence score. This highlights the tradeoff between regression quality and confidence score quality. These results correspond to HD-EPIC-S2 for CG-DETR. .

Table E. Performance of alternative confidence calibration mechanisms for CG-DETR on HD-EPIC-S2.

Variant	R_m			mAP_m			active
	@0.1	@0.3	@0.5	@0.1	@0.3	@0.5	
base	24.71	15.52	7.89	32.15	20.1	10.29	3.64 ± 1.18
+actionness[10]	26.04	16.05	8.11	32.26	19.96	10.23	3.79 ± 0.98
+SFL[7]	21.83	11.22	4.82	27.54	14.48	6.31	3.05 ± 0.99
-SA+QD (ours)	26.17	17.00	9.40	35.38	23.39	13.04	6.43 ± 2.16

As discussed in the main text, even though our proposed architectural modifications generally improve in both single and multi-moment queries, we observe a more prominent improvement in the latter. This aligns with the core motivation of the proposed modifications that specifically aim to activate more decoder queries, thus being able to detect more GT moments.

Table F. Dissection of the performance between "single" and "multi" for CG-DETR on HD-EPIC-S1/S2/S3

Input data	Model	Split	R_m			mAP_m		
			@0.1	@0.3	@0.5	@0.1	@0.3	@0.5
S1	base	single	31.47	20.09	10.19	39.29	25.02	12.68
		multi	22.90	13.73	6.77	29.97	18.46	9.57
	-SA+QD	single	31.61	21.39	12.25	41.72	28.23	16.25
		multi	26.18	16.16	8.11	35.66	22.89	12.12
S2	base	single	26.62	16.90	8.50	34.55	21.56	10.87
		multi	22.06	13.59	7.01	28.78	18.04	9.46
	-SA+QD	single	27.02	18.14	10.34	36.71	24.66	13.86
		multi	24.78	15.29	8.05	33.36	21.53	11.84
S3	base	single	14.66	8.96	4.06	21.87	13.03	6.11
		multi	8.63	3.67	1.61	15.47	6.99	3.01
	-SA+QD	single	15.64	10.36	6.16	23.23	15.31	8.88
		multi	9.26	5.47	2.67	15.71	9.39	4.61

Table G. Dissection of the performance between "single" and "multi" for CG-DETR on YC2-S.

Input data	Model	Split	R_m			mAP_m		
			@0.1	@0.3	@0.5	@0.1	@0.3	@0.5
S	base	single	39.43	28.70	17.54	50.62	36.31	22.06
		multi	24.22	15.89	8.40	33.58	22.77	12.12
	-SA+QD	single	39.26	29.47	17.88	51.83	39.44	24.75
		multi	25.82	16.08	8.48	36.18	24.81	13.8

Table H. Dissection of the performance between "single" and "multi" for CG-DETR on ANC-S.

Input data	Model	Split	R_m			mAP_m		
			@0.1	@0.3	@0.5	@0.1	@0.3	@0.5
S	base	single	60.63	39.87	24.59	72.39	54.65	37.12
		multi	60.26	42.58	24.46	71.72	55.32	35.10
	-SA+QD	single	65.03	42.55	26.00	75.0	56.61	38.45
		multi	61.72	44.16	24.70	72.42	56.24	35.21

H. Quantifying the language vs multi-moment gap

In this section, we present the full results (see Tab. I) corresponding to Fig. 4 of the main text. These results compare performance across the two given setups:$(\mathcal{D}^{captions}_{single}, \mathcal{D}^{search}_{single})$ and $(\mathcal{D}^{captions}_{multi}, \mathcal{D}^{search}_{multi})$. That is, for a given level of specificity—e.g., HD-EPIC-S2— we first identify the subset of single-moment search queries $\mathcal{D}^{search}_{single}$. This is, the search queries that even after under-specification still map to a single GT. Then we construct the subset $\mathcal{D}^{captions}_{single}$ corresponding to the very same instances but with the caption-based query. We repeat this process for the search queries that map to multiple moments, forming the remaining subsets $\mathcal{D}^{search}_{multi}$ and $\mathcal{D}^{captions}_{multi}$.

The purpose of this construction is to evaluate the same subset—single and multi-moment instances, respectively—varying only the specificity of the textual queries. Evaluating $\mathcal{D}^{*}_{single}$ isolates the effect of the language gap, since these subsets do not contain any instance that maps to multiple moments. In contrast, $\mathcal{D}^{*}_{multi}$ evaluates the compounded effect of the language and the multi-moment gap as we are able to contrast the performance from standard caption-based queries to their corresponding search queries that contain a considerable language gap and that necessarily map to multiple moments.

Table I. Extended results of the language vs multi-moment gap analysis from Sec. 4.

Dataset	Split	Specificity	R_m @0.1	R_m @0.3	R_m @0.5	mAP_m @0.1	mAP_m @0.3	mAP_m @0.5
HD-EPIC-S1	single	Orig.	38.53	24.81	13.04	46.04	29.74	15.74
		S	31.47	20.09	10.19	39.29	25.02	12.68
	multi	Orig.	25.97	15.03	7.74	35.75	21.21	11.2
		S	22.90	13.73	6.77	29.97	18.46	9.57
HD-EPIC-S2	single	Orig.	38.79	25.32	13.26	46.26	30.25	16.03
		S	26.18	16.90	8.50	34.55	21.56	10.87
	multi	Orig.	28.13	16.26	8.49	37.5	22.18	11.68
		S	22.06	13.59	7.01	28.78	18.04	9.46
HD-EPIC-S3	single	Orig.	32.13	20.78	11.72	39.47	25.83	14.22
		S	14.66	8.96	4.06	21.87	13.03	6.11
	multi	Orig.	33.67	20.89	10.57	42.13	26.29	13.57
		S	8.63	3.67	1.61	15.47	6.99	3.01
YC2-S	single	Orig.	60.30	46.98	35.67	67.22	52.21	38.94
		S	39.43	28.70	17.54	50.62	36.31	22.06
	multi	Orig.	64.70	52.41	38.75	70.48	57.07	41.99
		S	24.22	15.89	8.40	33.58	22.77	12.12
ANC-S	single	Orig.	74.83	59.51	43.82	81.67	68.71	52.69
		S	60.63	39.87	24.59	72.39	54.65	37.12
	multi	Orig.	76.75	61.00	44.49	83.51	70.6	54.00
		S	60.26	42.58	24.46	71.72	55.32	35.21

I. Extended ablations

Below we complement the ablation studies shown in Sec. 5 of the main text since because of space constraints, these include only the average R_m and mAP_m scores. Find below the corresponding results for each of the considered IOU scores. Concretely, Tab. J shows the complete results of the main ablations regarding alternative methods to increase the number of active decoder queries. Table L presents the results of the ablation that studies the impact of the 1-to-1 matching strategy as a diversity promoting mechanism in our proposed (-SA+QD). Table M further details the results of the ablation that investigates the impact of the QD dropout rate. Finally, Tab. N extends the results from the main text regarding the impact of increasing the total number of decoder queries.

Given that these findings remain consistent across benchmarks, for brevity we do not show the results for additional benchmarks.

Table J. Extended ablation of alternative methods to increase the number of active decoder queries, evaluated with CG-DETR on HD-EPIC-S2.

Variant	R_m 0.1	R_m 0.3	R_m 0.5	mAP_m 0.1	mAP_m 0.3	mAP_m 0.5
base	24.71	15.52	7.89	32.15	20.1	10.29
+ 1-to-5 matching [9]	24.00	15.11	7.97	29.35	18.36	9.6
+ 1-to-k matching [9]	18.31	8.80	4.12	18.47	8.87	4.16
+group_matching [2]	24.33	15.85	8.86	31.21	20.12	11.14
+hybrid matching [6]	23.98	15.19	7.75	30.53	19.07	9.71
+ms matching [14]	24.13	15.30	8.02	31.72	20.35	10.79
+data augmentation	20.96	13.14	7.31	31.64	20.52	11.55
-SA+QD (ours)	26.17	17.00	9.40	35.38	23.39	13.04

Table K. Extended ablation of the effect of the 1-to-1 matching strategy in promoting diversity across decoder queries, for CG-DETR evaluated on HD-EPIC-S2

Variant	R_m 0.1	R_m 0.3	R_m 0.5	mAP_m 0.1	mAP_m 0.3	mAP_m 0.5
-SA+QD (ours)	26.17	17.00	9.40	35.38	23.39	13.04
+ 1-to-k matcher [9]	17.88	9.18	4.10	17.89	9.19	4.11
+group matcher [2]	25.72	17.17	9.02	35.36	23.37	12.41
+hybrid matcher [6]	26.59	17.60	9.55	35.79	24.13	13.24

Table L. Extended ablation of the impact of each of the proposed architectural modifications, for CG-DETR evaluated on HD-EPIC-S2.

-SA	+QD	R_m 0.1	R_m 0.3	R_m 0.5	mAP_m 0.1	mAP_m 0.3	mAP_m 0.5
		24.71	15.52	7.89	32.15	20.1	10.29
✓		23.97	14.73	7.25	32.17	20.58	10.33
	✓	24.45	16.24	8.83	31.58	21.07	11.66
✓	✓	26.17	17.00	9.40	35.38	23.39	13.04

Table M. Extended ablation of the effect of the QD dropout rate, for CG-DETR evaluated on HD-EPIC-S2.

k	R_m			mAP_m		
	0.1	**0.3**	**0.5**	**0.1**	**0.3**	**0.5**
0.00	24.71	15.52	7.89	32.15	20.1	10.29
0.25	26.17	17.00	9.40	35.38	23.39	13.04
0.50	1.82	0.86	0.31	6.72	3.49	1.32

Table N. Extended ablation (corresponding to Fig. 8 of the main text) that shows the impact of the overall number of decoder queries for the CG-DETR and our proposed (-SA + QD), for HD-EPIC-S2.

Model	**# queries**	R_m			mAP_m			**# active**
		0.1	0.3	0.5	0.1	0.3	0.5	
base	5	22.41	14.12	6.94	27.78	17.46	8.71	2.87 ± 0.70
	10	23.60	15.39	7.65	29.4	19.05	9.59	3.32 ± 1.06
	20	24.71	15.52	7.89	32.15	20.1	10.29	3.52 ± 1.05
	30	21.88	15.00	7.86	28.83	19.65	10.39	3.98 ± 1.28
	50	11.66	6.10	2.74	18.12	10.02	4.58	3.66 ± 1.10
-SA + QD (ours)	5	15.88	7.89	3.31	20.8	10.3	4.35	2.64 ± 1.52
	10	25.97	17.37	9.12	34.02	22.57	12.04	5.20 ± 1.35
	20	26.17	17.00	9.40	35.38	23.39	13.04	6.73 ± 1.86
	30	20.71	10.70	4.91	29.64	15.71	7.23	6.78 ± 1.47
	50	16.54	7.00	2.58	25.79	12.42	5.08	6.49 ± 1.48

J. Qualitative results of the generated search queries

Here we show various qualitative examples of the search queries generated by our proposed search-query pipeline. More concretely, for each of the benchmarks, we first showcase various examples of the under-specifications resulting from the original caption-based queries. Then, we include various instances of final search queries, showing all the original captions that resulting from the search-query pipeline, end up mapping to the same representative search query.

J.1. HD-EPIC

As explained in Sec. C, thanks the great detail of the captions of this dataset, we were able to extract 3 different levels of under-specified search queries—S1,S2 and S3. Figure M further depicts the effect of these under-specifications by visualizing the features similarities of the under-specified search queries with respect to the original caption-based ones.

In Fig. O, moreover, we present multiple qualitative examples of how a caption-based query progressively under-specifies in each of the proposed levels of specificity. This is further shown in Fig. P, Fig. Q and Fig. R where we show various final search queries, including the final search query and the multiple original captions that match it.

J.2. YC2-S

Below we repeat the same analysis for YC2-S benchmark. Concretely, find in Fig. N the histogram of the features similarities between the original caption-based queries and the corresponding search queries. Additionally, Fig. S provides various qualitative examples of under-specifications of original captions from YC2 [16] into their corresponding search queries, while Fig. T shows various final search queries.

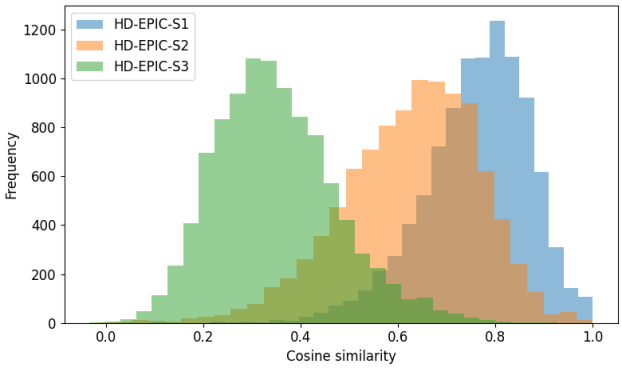

Figure M. Histogram of the feature similarities of each of the test set of each of the levels of under-specification HD-EPIC-S1/S2/S3 with respect to the original caption-based queries from HD-EPIC

Table O. Qualitative results of the under-specified search queries for HD-EPIC-S1/S2/S3.

Caption: *"Holding the ball of chicken and potato mixture in my left hand while I get some flour in my right hand so that I can sprinkle over the ball"*

↓

S1: *"Hold the mixture and get some flour to sprinkle over it"*

↓

S2: *"Sprinkle flour on the mixture"*

↓

S3: *"Prepare food"*

Caption: *"Pick up tissue from inside the plate on the countertop using the right hand ."*

↓

S1: *"Pick up tissue from the plate"*

↓

S2: *"Pick up tissue"*

↓

S3: *"Pick up something"*

J.3. ANC-S

Finally, we perform the same analysis for ANC-S benchmark. Concretely, find in Fig. O the histogram of the features similarities between the original caption-based queries and the corresponding search queries. Additionally, Fig. U provides various qualitative examples of under-specifications of original captions from ANC [8] into their corresponding search queries, while Fig. V shows various final search queries.

Table P. Example of search queries and the captions that match it for HD-EPIC-S1.

Representative: *"Tear the lettuce and place it on the plate"*
↓
Caption1: *"Tear the lettuce leaves in half again and place onto the plate that is on top of the weighing scale ."*
Caption2: *"Tear the lettuce leaves in half again and put the pieces onto the plate ."*
Caption3: *"Use both hands to tear the lettuce leaves in half again and place the pieces onto the plate ."*
Caption4: *"Use both hands to tear the lettuce leaves in half ."*
Caption5: *"Use both hands to tear the lettuce leaf in half"*
Caption6: *"Use the left hand to put the lettuce leaves into the bowl and then use both hands to tear the lettuce leaves in half again ."*
Caption7: *"Use both hands to tear the lettuce leaves in half."*
Caption8: *"Use both hands to tear the lettuce leaves in half."*
Caption9: *"Use both hands to grip the lettuce leaves and tear them in half."*
Caption10: *"Use both hands to tear the lettuce leaves in half again."*
Caption11: *"Use both hands to tear the lettuce leaves in half and the right hand to place the leaves into the plate that is on the weighing scale."*
Representative: *"Sprinkle flour over the chicken ball"*
↓
Caption1: *"Using my right hand to retrieve some flour from the bowl of flour while my left hand holds the chicken ball in place so that I can sprinkle some flour over it ."*
Caption2: *"Bringing my right hand closer to the potato chicken ball as I move it around in my left hand and sprinkle flour over it with my right hand ."*
Caption3: *"Using my right hand to sprinkle flour over the potato chicken ball"*
Caption4: *"Using my right hand to scoop some flour to sprinkle over the potato chicken ball held in my left hand so that I can use up more flour ."*
Caption5: *"Placing the potato chicken ball in my left hand using my right hand so that I can sprinkle more flour over it ."*
Caption6: *"Sprinkling flour over the potato chicken ball using my right hand while holding potato chicken ball in my left hand"*
Caption7: *"Sprinkling flour over the potato chicken all in my left hand ."*

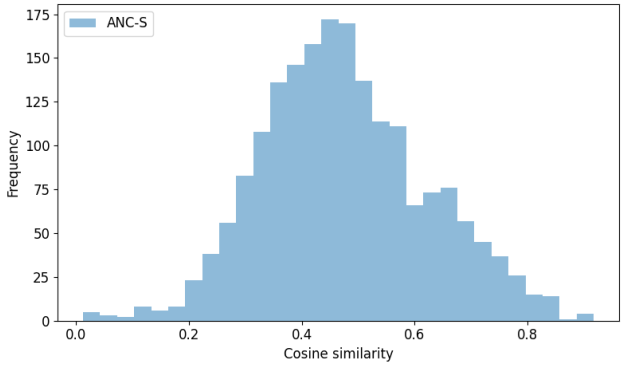

Figure N. Histogram of the feature similarities of the test set of YC2-S with respect to the original caption-based queries from YC2

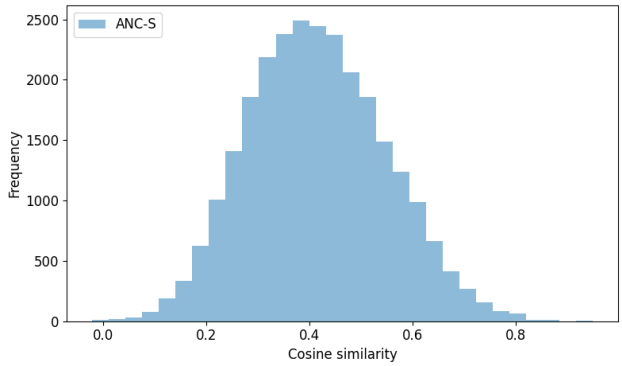

Figure O. Histogram of the feature similarities of each of the test set of ANC-S with respect to the original caption-based queries from ANC

References

[1] Zhuo Cao, Bingqing Zhang, Heming Du, Xin Yu, Xue Li, and Sen Wang. Flashvtg: Feature layering and adaptive score handling network for video temporal grounding. In *2025*

Table Q. Example of search queries and the captions that match it for HD-EPIC-S2.

Representative: *"Close a cupboard"*

↓

Caption1: *"With my right hand , close the bottom cupboard ."*
Caption2: *"With my right hand , close the cupboard ."*
Caption3: *"Close the cupboard by pushing it with my left hand ."*
Caption4: *"Having realized this is the wrong cupboard , close it again by pushing it with my left hand ."*
Caption5: *"I close the cupboard by pushing it with my right hand ."*
Caption6: *"Close the cupboard with my right hand ."*
Caption7: *"Close the cupboard by pushing it with my left hand ."*
Caption8: *"With my right hand , close the cupboard by pushing it ."*
Caption9: *"Not finding what I was looking for , close again the cupboard with my left hand ."*
Caption10: *"With my left hand , close the cupboard by pulling it towards me ."*
Caption11: *"With my right hand , close the smaller cupboard on the right hand side ."*
Caption12: *"With my right hand , close the cupboard by pulling the cupboard towards me ."*
Caption13: *"Close the cupboard using my left hand ."*
Caption14: *"With my right hand , close the cupboard by pulling the cupboard ."*
Caption15: *"Using my left hand , close the other larger cupboard ."*
Caption16: *"Close the small cupboard using my right hand ."*
Caption17: *"With my right hand , close the top cupboard ."*
Caption18: *"With my right hand , close the bottom cupboard ."*
Caption19: *"With my left hand , close the cupboard in the corner next to the dishwasher by pushing the cupboard ."*

Representative: *"Open a lid"*

↓

Caption1: *"Open the lid of the trash bin by flipping it ."*
Caption2: *"Open the trash bin 's lid ."*
Caption3: *"Open the lid of the trash bin . This action occurs off the screen ."*
Caption4: *"Open the lid of the food waste bin ."*

IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 9226–9236. IEEE, 2025.

[2] 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.

[3] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Jimmy Lin. Ms marco: Benchmarking ranking models in the large-data regime. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 1566–1576, 2021.

[4] 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.

[5] Kevin Flanagan, Dima Damen, and Michael Wray. Moment of untruth: Dealing with negative queries in video moment retrieval. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5336–5345. IEEE, 2025.

[6] 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.

[7] Viacheslav Komisarenko and Meelis Kull. Improving calibration by relating focal loss, temperature scaling, and properness. *arXiv preprint arXiv:2408.11598*, 2024.

[8] 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.

[9] 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.

[10] Xiaolong Liu, Qimeng Wang, Yao Hu, Xu Tang, Shiwei Zhang, Song Bai, and Xiang Bai. End-to-end temporal ac-

Table R. Example of search queries and the captions that match it for HD-EPIC-S3.

Representative: *"Open a book"*

↓

Caption1: *"Push down in the center of the recipe book along the spine to try and make sure it stays open onto the countertop."*
Caption2: *"Pick up the recipe book using the right hand and flipping it over so that the recipe can be seen."*

Representative: *"Clean a bowl."*

↓

Caption1: *"Realise there is some old food or dirt on the bottom of the bowl."*
Caption2: *"Try to scratch away the food or dirt to see if it's on the bottom of the bowl or inside the bowl using the right hand."*
Caption3: *"Clean the bottom of the bowl using the kitchen roll in the right hand whilst holding the bowl down on the left hand."*
Caption4: *"Grab hold of the sponge using the right hand and continue to clean the inside of the large mixing bowl."*
Caption5: *"Submerge the bowl into the water and clean the inside of the bowl using the sponge in the right hand, paying close attention to get rid of any food that has been stuck onto the bowl."*
Caption6: *"Rotate the bowl over to clean the inside once more."*

Table S. Qualitative results of the under-specified search queries for YC2-S.

Caption: *"Add strained tomato puree to the blond roux and stir the mixture continuously"*

↓

S: *"Add ingredient."*

Caption: *"Add salt to the pan and mix"*

↓

S: *"Season food."*

tion detection with transformer. *IEEE Transactions on Image Processing*, 31:5427–5441, 2022.

[11] 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.

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

[13] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer, 2024.

[14] 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.

[15] 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.

[16] 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.

Table T. Example of search queries and the captions that match it for YC2-S.

Representative: *"Add ingredients."*   ↓   **Caption1:** *"Crack one egg into a bowl"*   **Caption2:** *"Add one table spoon of oil salt and cayenne pepper and baking powder and beat"*   **Caption3:** *"Add one cup of beer and mix"*   **Caption4:** *"Add one quarter cup of corn meal and one cup of flour"*   **Caption5:** *"Add onions into batter and drop into hot oil"*
Representative: *"Cook potatoes."*   ↓   **Caption1:** *"Spread rock salt on a baking tray place potatoes on it and draw few spikes."*   **Caption2:** *"Pierce the knife inside the potatoes and find if the potatoes are cooked properly."*   **Caption3:** *"Cut the cooked potatoes in half and scoop the flesh and put it in a bowl."*   **Caption4:** *"Keep the mashed potatoes on the flame and mix adding butter until it is mix well and the bottom of the pan becomes shiny."*

Table U. Qualitative results of the under-specified search queries for ANC-S.

Caption: *"The man is washing a side of the car."*   ↓   **S:** *'A person performs a task on a vehicle."*
Caption: *"The person then moves back and fourth on the machine while rowing his arms back and fourth."*   ↓   **S:** *"A person uses a machine."*

Table V. Example of search queries and the captions that match it for ANC-S.

Representative: *"People play a game."*   ↓   **Caption1:** *"A man and woman are shown standing on a tennis court passing a ball back and fourth."*   **Caption2:** *"The players volley and the play is pressed to the back of the court hitting long shots."*   **Caption3:** *"The players volley and the birdie is hit out of view and the player retrieves it then serves."*   **Caption4:** *"The players have a long volley until the play in the foreground misses and the birdie lands at her feet."*   **Caption5:** *"People play games of badminton on indoor courts."*   **Caption6:** *"Behind them are another group of people playing and the man and woman continuously pass the ball back ad fourth to one another."*
Representative: People perform household chores.   ↓   **Caption1:** *"A young girl and boy are washing dishes in a kitchen."*   **Caption2:** *"They are doing the dishes in the sink."*   **Caption3:** *"the mother enters and adds some dishes from the rack back into the sink to be rinsed again and shows the boy what was wrong with the pot."*