Supplementary - Moment of Truth: Dealing with Negative Queries in Video Moment Retrieval

Kevin Flanagan
University of Bristol
kevin.flanagan@bristol.ac.uk

Dima Damen
University of Bristol
dima.damen@bristol.ac.uk

Michael Wray
University of Bristol
michael.wray@bristol.ac.uk

In this supplementary, we provide more information about the dataset creation in Sec. 1, describing the process of generating the out-of-domain negative queries using LLMs and demonstrating our prompts and categories as well as details about the Negative-Aware Video Moment Retrieval dataset. We display full results for SVMs trained on output saliency scores for all three models (UniVTG [3], CG-DETR [4], QD-DETR [5]) in Sec. 2. We expand on the adjustments made to the losses for UniVTG-NA and detail the QD-DETR and CG-DETR implementations in Sec. 3. We demonstrate the out-of-domain generalisability of UniVTG-NA in Sec. 4. We further motivate the need for negative-aware methods in video moment retrieval by displaying results with negative queries from UniMD [6] in Sec. 5. Finally, we show more qualitative results from UniVTG-NA on QVHighights and Charades-STA in Sec. 6.

1. Dataset Information

1.1. Out-of-Domain Negative Query Generation

As mentioned in the main paper, the out-of-domain negative query sentences were generated using a large language model (LLM). Four broad scenarios were used as query topics, these were "competitive sport", "animal behaviour", "physics laboratory", and "mathematics class". In Table 1, the subtopics for each of these topics are listed. The prompt and specific LLM used for each topic is also displayed. Prompts were empirically chosen to ensure the quality and diversity of the generated sentences. For example, for the "animal behaviour" topic, it was found that using scientific names improved upon these aspects, hence most of the subtopics are scientific names. The four scenarios were chosen as they represent scenarios which are unlikely to be present within the QVHighlights and Charades-STA datasets, which cover news, vlogs and household actions. The choice of 4 broad scenarios helps to ensure that the OOD negatives remain OOD and do not accidentally produce false negatives. By using prompts specifically describing the actions as short, unique and varied, we are able to get a wider range of sentences without the language becoming too decorative. Sample sentences from each scenario are displayed in Table 2. We use the same set of OOD Negatives for both QVHighlights [2] and Charades-STA [1].

1.2. Negative Aware Dataset Details

Table 3 displays the number of positive and negative queries used during training/evaluation of the models. For Charades-STA, where there are fewer negative queries than positive for out-of-domain, the negative queries are assigned to multiple videos. This still produces a distinct signal as each video-sentence pair offers a different semantic relationship.

2. SVM Trained on Saliency Scores

We train an SVM on the outputted positive and negative query saliency scores from UniVTG [3], QD-DETR [5] and CG-DETR [4] for the QVHighlights and Charades-STA datasets. This is to quantify how separable positive and negative queries are when the relationship between them is modelled using saliency outputs from the base models, without any explicit training for negative rejection. Results are displayed in Table 4.

The SVM results on QVHighlights show high rejection accuracy at the cost of decreased R1@θ scores. In the case of QD-DETR, these are significantly decreased. The Charades-STA results show reasonable rejection accuracy at significant cost to the R1@θ scores for CG-DETR and QD-DETR, while UniVTG fails to achieve high rejection accuracy but has better R1@θ scores. Overall these results display the limitations of using the saliency outputs from the base models alone for combined moment retrieval and negative rejection, particularly on datasets without ground-truth saliency scores such as Charades-STA. It further motivates the need to train explicitly for negative rejection.

Topic	Competitive Sport		Animal Behaviour			Physics Laboratory	Mathematics Class
Model	Chat-GPT (GPT-4o)		Claude 3 Opus			Claude 3 Opus	Claude 3 Opus
Prompt	Generate X sentences describing actions in <subtopic>		Generate X unique and varied short sentences of visual actions carried out by <subtopic>			Generate X unique and varied sentences of visual actions carried out by a person working in a <subtopic> lab.	Generate X unique and varied sentences of visual actions carried out by a person working in a <subtopic> class.
Subtopics	american football	archery	accipitriformes	agnatha	alcidae	acoustics	algebra
	athletics field events	badminton	anatidae	anguilliformes	annelids	atmospheric physics	applied mathematics
	baseball	boxing	anura	big cats	bivalves	biophysics	calculus
	cricket	cycling	bovidae	camelid	canidae	chemical physics	combinatorics
	darts	fencing	cephalopods	cervidae	chelicerata	classical mechanics	computational maths
	field hockey	golf	chiroptera	chondrichthyes	cnidaria	condensed matter physics	geometry
	gymnastics	ice hockey	crocodilia	decapods	echinoderms	cosmology	graph theory
	ice skating	kickboxing	elasmobranchs	gastropods	giraffidae	electromagnetism	number theory
	lacrosse	rowing	hymenoptera	insects	lagomorphs	electronics	probability
	rugby	running	lepidoptera	lizards	marsupials	fluid dynamics	statistics
	skateboarding	skiing	monotremes	mustelids	osteichthyes	geophysics	
	snooker	snowboarding	pinnipeds	platyhelminthes	porifera	medical physics	
	soccer	squash	primates	proboscidea	ratites	optical physics	
	swimming	table tennis	rodents	serpentes	spheniscidae	particle physics	
	tennis	ultimate frisbee	stomatopods	strigiformes	suina	quantum mechanics	
	water polo		talpidae	testudines	urodela	thermodynamics	
			ursidae	wading birds			

Table 1. List of topics and subtopics used for out-of-domain negative generation, along with the prompts and LLMs used. X represents the number of sentences requested which varied from 50 to 100.

3. Model Details

3.1. UniVTG-NA

For UniVTG-NA, the input to the classification head is a direct sum of the indicator scores and saliency scores. *i.e.* $g_i = f_i + s_i$ where g_i is the classification head input at index i.

Loss Adaptations. We specify the adjustments made to the losses for the UniVTG-NA model from UniVTG [3]. Aside from the boundary prediction losses being set to 0 for the negative queries, the saliency losses are also adjusted. UniVTG uses a saliency loss $\mathcal{L}_s$ which is a weighted summation of inter-video and intra-video contrastive losses. It is not possible to use the contrastive saliency loss with negative queries. Therefore, for negative queries the saliency loss is defined as a loss applied directly on the cosine similarity between the video clip $\mathbf{v}_i$, and sentence features $\mathbf{S}$, with λ_s^- as the loss weighting.

$$\mathcal{L}_s^- = \lambda_s^- \cos(\mathbf{v}_i, \mathbf{S}) := \lambda_s^- \frac{\mathbf{v}_i^T \mathbf{S}}{\|\mathbf{v}_i\|_2 \|\mathbf{S}\|_2} \tag{1}$$

This is done as the saliency scores are computed via cosine similarity between sentence and video clip features for UniVTG, so achieves our principle of designing the saliency loss for negatives such that it pushes the saliency scores lower. Furthermore, for UniVTG-NA's foreground matching loss with negative queries, $\mathcal{L}_f^- = \mathcal{L}_f$ as no adjustments are made to the matching loss, which is a BCE loss on the individual indicator scores. This already achieves the aim of pushing the indicator scores lower.

3.2. QD-DETR-NA & CG-DETR

Negative-aware versions of QD-DETR and CG-DETR were also trained to evaluate the proposed method on other models. The details of the QD-DETR and CG-DETR implementation are as follows: Given the indicator score outputs $\{\tilde{f}_1, ..., \tilde{f}_M\}$ and saliency score outputs $\{\tilde{s}_1, ..., \tilde{s}_{L_v}\}$, the input to the classification head is a concatenation $g = \{\tilde{s}_1, ..., \tilde{s}_{L_v}, \tilde{f}_1, ..., \tilde{f}_M\} \in \mathbb{R}^{(L_v+M)}$, where L_v is the number of video clip features and M is the number of moment queries. This implementation is represented in Figure 1. This is chosen as opposed to a summation because QD-DETR/CG-DETR use moment queries to generate the moment candidates rather than just the text-attended video clip representations from the encoder. In this case, there is not a one-to-one correspondence with the saliency scores, *i.e.* $L_v \neq M$.

As with UniVTG, the boundary losses were set to 0 and the foreground matching loss was retained for the negative queries. For both methods, the saliency loss has three components, two of which are contrastive and are therefore not feasible for negative queries. The remaining loss works to reduce the negative query saliency scores, thus achieving the principle aim of the negative query saliency loss. Therefore it is retained as the sole loss for negative queries. It is shown for a saliency score output s_i with loss weighting λ_s^- below.

$$\mathcal{L}_s^- = \lambda_s^- (-\log(1 - s_i)) \tag{2}$$

Competitive Sport	Animal Behaviour
The outfielder throws to home plate	An osprey dives into the water, snatching a fish with its talons
The opponent hits a drop shot followed by a lob	A northern harrier glides low over a meadow, searching for small mammals.
The striker heads the ball into the net	An ovambo sparrowhawk sits near its nest, guarding its eggs.
The punter pins the opposing team deep in their own territory with a well-placed kick	A slender-billed kite hunts for insects over an African grassland
The player executes a deceptive backhand drop shot	A white-backed vulture strips meat from a carcass with its strong beak
The opponent flicks a shuttlecock deep into the backcourt	A lamprey swam in a figure-eight pattern, leaving pheromone trails for potential mates
The goalie makes a sprawling save	A group of lamprey larvae anchored themselves to rocks, facing into the current
The player switches to a colored ball after potting all reds	A bronze eel lay coiled on the seafloor, its coppery scales gleaming
The opponent covers up, absorbing the blows	A topaz eel darted through a school of yellow tang, its golden body mirroring their color
The center offloads the ball to a teammate before being tackled	A bootlace worm tangles itself around a piece of driftwood
The flanker disrupts the opposing team's maul, forcing a turnover	A red-eyed tree frog clings to a leaf with its sticky toe pads
The rowers maintain their balance as the boat rocks gently on the water	The Amazon milk frog inflated its body, trying to appear larger
The rowers return to the dock and disembark from the boat	The jaguar's powerful jaws crushed the turtle's shell
The archer aims downrange, focusing on the target	A long clam extends its siphons, drawing in water to filter out food particles
The skater lands a double axel with precision	A kudu reached up to browse on acacia tree leaves
The swimmer's streamline position reduces resistance through the water	A Pale fox kit playfully wrestles with its sibling outside their den
Skiers maintain a tight tuck to minimize drag	A Cape fox, known for its nocturnal habits, emerges from its den at dusk to begin hunting
The athlete lands on the mat on the other side of the bar	The moose browsed on the tender bark of a young tree, stripping it with its teeth
The skater executes a jump combination, linking jumps of different rotations	A crab spider ambushed a bee from its hiding spot in a flower
The gymnast tumbles with precision on the floor exercise mat	A moon coral's large, rounded polyps resemble a cluster of full moons
Physics Laboratory	**Mathematics Class**
The researcher measured the sound absorption coefficient of the new acoustic material	They create a flow diagram to show the steps in the algorithm
The acoustician measured the sound reduction index of the window using a pink noise generator	He draws a box-and-whisker plot to compare the distributions of different data sets
He used a sound intensity probe to measure the sound power of the jet engine	They create a Venn diagram to find the probability of the union of two events
The researcher uses a ceilometer to determine the height of the cloud base	She uses the separation of variables technique to solve the partial differential equation
With a steady hand, the chemist uses a capillary tube to load the viscous ionic liquid into the rheometer for flow behavior studies	He arranges a set of numbered tiles to illustrate the concept of permutations with repetition
The graduate student intently studies the XPS spectrum, identifying the chemical states of the elements present on the catalyst surface	With a critical eye, she examined the partial dependence plots, assessing the impact of individual features on the model
She recorded the data from the oscilloscope in her lab notebook	He labels each vertex with a unique letter, making it easier to refer to specific nodes
The scientist replaced the filament in the electron gun	She shades a vertex to indicate it has been visited during a graph traversal
He carefully positioned the sample in the center of the split-coil magnet	He draws a graph with a minimum spanning tree, a subgraph connecting all vertices with the minimum total edge weight
The scientist adjusted the settings on the surface plasmon resonance (SPR) instrument	He shades the area representing the union of two probability events
The cosmologist carefully positioned the spectrograph, ready to analyze the light from a distant supernova	The statistician carefully folded the large printed graph, ensuring the creases were sharp and the edges aligned
He studies the flow patterns in a porous medium using magnetic resonance imaging	The analyst used a highlighter to trace the trend line on the time series plot
The researcher measures the thermal conductivity of a rock sample using a divided bar apparatus	The statistician used a chalk line to draw a perfectly straight line on the chalkboard, representing the regression equation
The geophysicist uses a Schmidt hammer to test the strength of a rock outcrop	He leans forward, listening intently to his colleague's explanation of a new mathematical technique
The physicist calibrated the radionuclide calibrator for accurate activity measurements of radiopharmaceuticals	The mathematician creates a Pascal's triangle, highlighting the connection between combinatorics and binomial coefficients
The scientist adjusted the position of the camera to capture the desired image	She writes out the formula for calculating the number of combinations of n objects taken r at a time
He measured the wavelength of the light using a spectrometer	He creates a matrix to represent the adjacency relationships in a combinatorial graph
She calculates the probability of a defective product using quality control data	He arranges a set of dominoes in different configurations, exploring the number of possible tilings
She adjusts the phase shifter to control the interference between the microwave signals	The mathematician draws a tree diagram to illustrate the Collatz conjecture
The scientist uses a laser thermometer to measure the surface temperature of the material	She writes out a proof using mathematical induction, establishing a pattern
Musician Performance	
The accordionist's fingers danced across the keys, effortlessly transitioning between notes	
The man blows into the blowpipe to fill the bag with air	
The banjo player's hands moved in a blur, creating an intricate fingerpicking pattern	
He muted the strings with his palm, creating a staccato effect	
The bongo player's hands alternate between drums	
The cellist leans into the instrument, conveying the emotion of the piece through their posture	
She brushes the snare drum lightly, creating a soft, sizzling sound	
Their cheeks puff out as they blow into the mouthpiece	
She plays the guitar while sitting on a stool	
He tapped his fingers on the fretboard, creating a percussive rhythm	
He alternates between blowing and drawing on the harmonica, creating a dynamic sound	
With closed eyes, the musician swayed gently as they strummed the harp's delicate strings	
She gently presses the white keys with her fingertips	
She places her feet on the pedals and her hands on the keys	
She smiled at the audience, her saxophone gleaming under the stage lights as she played a upbeat tune	
He slides his left hand along the strings to change the pitch	
They play a glissando by sliding their finger across the keys	
They keep their hands steady for a long, sustained note	
He tilts the trombone up for a high note	
The musician's eyes darted between the sheet music and his fingers, ensuring he played each note correctly	

Table 2. Example sentences from each OOD topic.

	Train			Test		
	Positive	In-Domain Negative	Out-of-Domain Negative	Positive	In-Domain Negative	Out-of-Domain Negative
QVHighlights [2]	7218	7218	7230	1550	1550	1550
Charades-STA [1]	12404	12404	7230	3720	3720	1550

Table 3. Numbers of positive and negative queries used for QVHighlights and Charades-STA.

3.3. Implementation Details

UniVTG For QVHighlights, we use loss weightings of $\lambda^+ = 1$, $\lambda^-_{ID} = 0.1$, $\lambda^-_{OOD} = 0.1$, and $\lambda_p = 1$, while for Charades-STA, we adjust $\lambda^-_{ID} = 0.5$, $\lambda^-_{OOD} = 0.5$. The remaining loss weightings are retained from QVHighlights and Charades-STA training defaults in UniVTG. For negative queries the cosine similarity loss weighting λ^-_s is set equal to the intra video saliency loss weighting.

QD-DETR Loss weightings of $\lambda^+ = 1$, $\lambda^-_{ID} = 0.05$, $\lambda^-_{OOD} = 0.05$, and $\lambda_p = 1$ are used for both QVHighlights and Charades-STA. For QVHighlights, $\lambda^-_s = 1$ while for Charades-STA, $\lambda^-_s = 4$.

CG-DETR The same weightings are used as in QD-DETR except $\lambda^-_{ID} = 0.1$, $\lambda^-_{OOD} = 0.1$ for both datasets. All other loss weightings retain their default values.

4. OOD Generalisability

To test the generalisability of the negative-aware approach for OOD query sentences, we test the UniVTG-NA model on OOD sentences from another scenario on which the model has not been trained. This scenario is 'musician performances' (see sample sentences in Table 2). The re-

Method	QVHighlights				Charades-STA			
			Rejection Acc. (%)				Rejection Acc. (%)	
	R1@0.5	R1@0.7	ID	OOD	R1@0.5	R1@0.7	ID	OOD
UniVTG [3] SVM	63.48 (-3.87)	49.87 (-2.78)	94.77	97.74	53.47 (-6.75)	33.49 (-5.06)	35.89	50.40
CG-DETR [4] SVM	62.84 (-4.26)	50.26 (-3.29)	95.55	95.41	43.79 (-13.74)	28.44 (-7.23)	82.90	74.52
QD-DETR [5] SVM	48.26 (-13.74)	37.42 (-8.84)	96.32	95.09	46.67 (-12.44)	30.05 (-6.70)	76.91	81.59

Table 4. Results of training an SVM on top of the saliency score outputs of UniVTG, CG-DETR and QD-DETR.

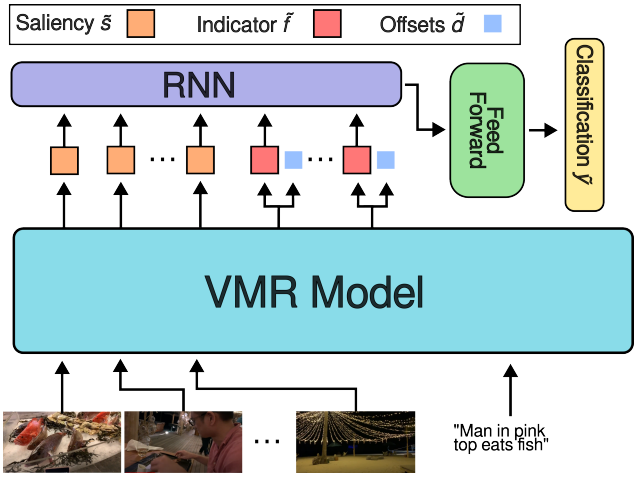

Figure 1. The classification head for QD-DETR-NA and CG-DETR-NA takes as input a concatenation of saliency and indicator scores, which are then passed through a recurrent layer and a feed forward layer before producing a single value output for classification.

jection accuracy results are shown in Table 5. The rejection accuracy remains high for both datasets, demonstrating that the model is capable of generalising to other OOD scenarios.

Method	Rejection Acc. (%)	
	QVHighlights	Charades-STA
UniVTG-NA	99.8	93.8

Table 5. Rejection accuracy results for UniVTG-NA on the unseen OOD category of 'musician performance'.

5. UniMD

To further motivate the need for negative-aware training for the task of Negative-Aware Video Moment Retrieval, we investigate the output produced by UniMD [6], a recent SOTA method which only produces indicator scores with no saliency scores. We plot histograms of the output scores for positive and in-domain negative sentences for Charades-STA and ActivityNet-Captions, as in Figure 2. There is significant overlap between the positive and negative distributions which shows that the model is not designed to handle negative rejection. This further motivates the need for models which are specifically trained to carry out negative rejection alongside moment retrieval.

6. Qualitative Results

We provide further qualitative results from UniVTG-NA on the QVHighlights and Charades-STA datasets in Figure 3 & 4. The model frequently successfully localises the positive sentences and rejects the negative sentences. Failure cases are included in the bottom right of each set of examples. The failure case in Figure 3 is a case of UniVTG-NA rejecting a positive sentence, while in Figure 4 the model fails to reject an ID negative sentence.

References

[1] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, 2017. 1, 3

[2] Jei Lei, Tamara L. Berg, and Bansal. Qvhighlights: Detecting moments and highlights in videos via natural language queries. In *NeurIPS*, 2021. 1, 3

[3] Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. Univtg: Towards unified video-language temporal grounding. In *ICCV*, 2023. 1, 2, 4

[4] WonJun Moon, Sangeek Hyun, SuBeen Lee, and Jae-Pil Heo. Correlation-guided query-dependency calibration in video representation learning for temporal grounding. *CoRR abs/2311.08835*, 2024. 1, 4

[5] WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *CVPR*, 2023. 1, 4

[6] Yingsen Zeng, Yujie Zhong, Chengjian Feng, and Lin Ma. Unimd: Towards unifying moment retrieval and temporal action detection. *ECCV*, 2024. 1, 4

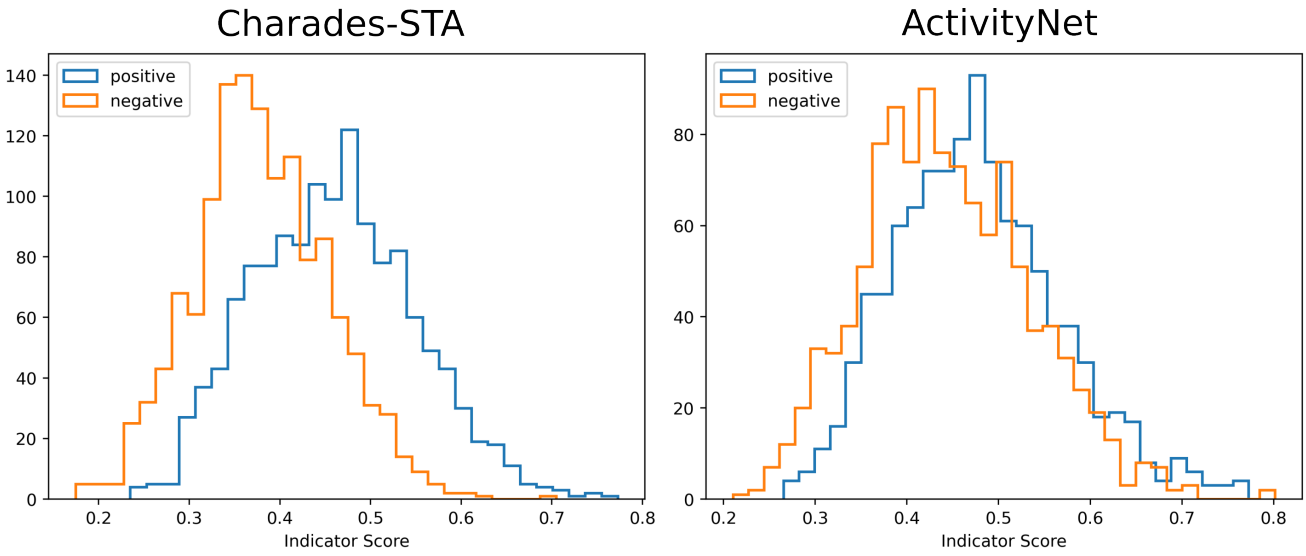

Figure 2. Histograms of prediction (indicator) scores for positive and in-domain negative queries produced by the UniMD model.

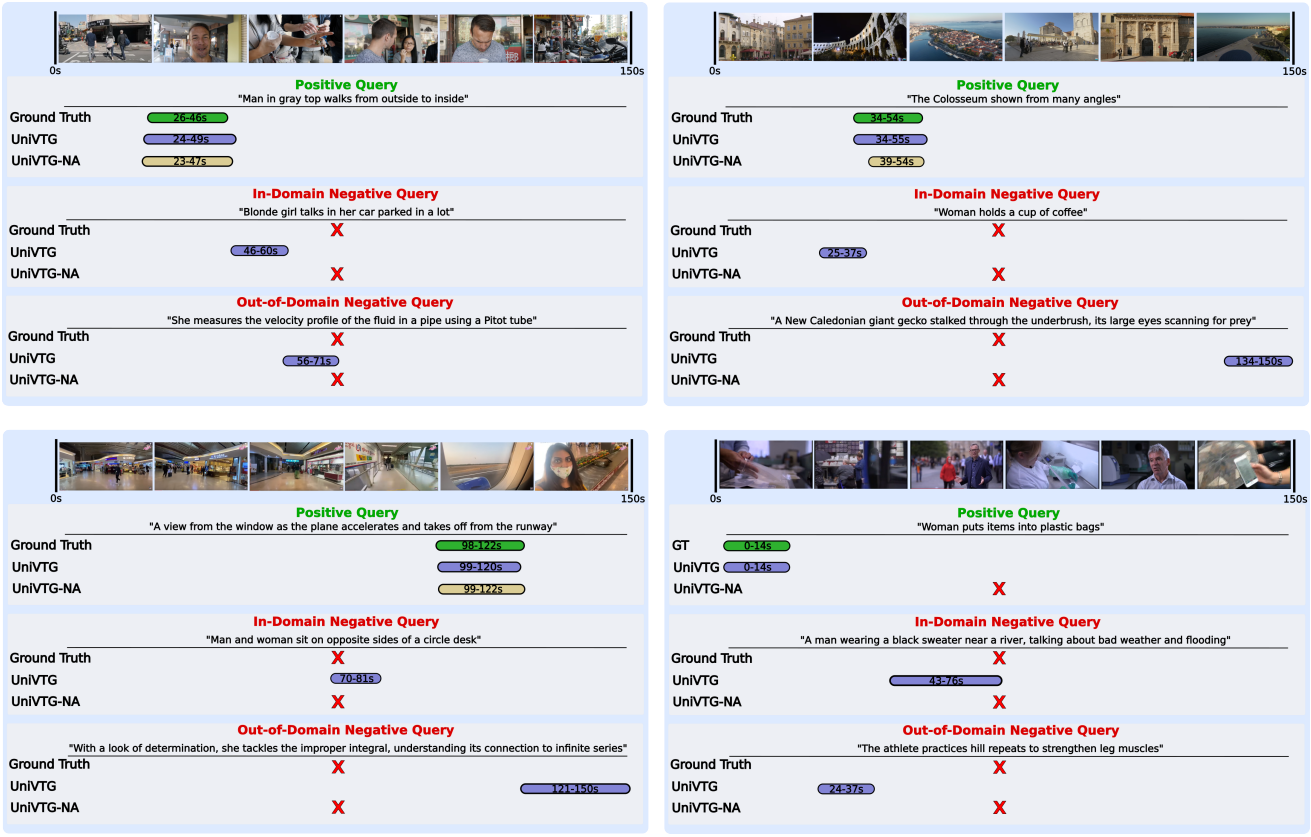

Figure 3. Qualitative results from UniVTG-NA on QVHighlights.

Figure 4. Qualitative results from UniVTG-NA on Charades-STA.