

Supplementary Materials: Let Me Finish My Sentence

Anonymous Authors

QUERY: A group of people stand around a table and take turns running their hands around the base of a mound of dirt on the table.

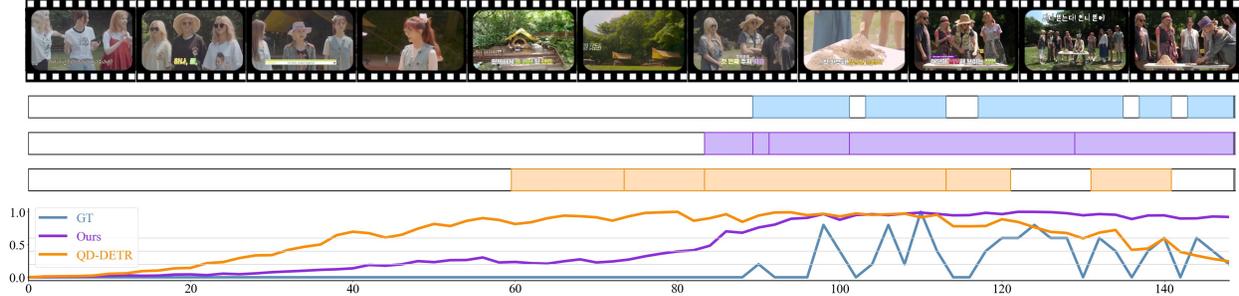


Figure 1: Comparative prediction results on the QVHighlights validation set for longer text queries, demonstrating our method’s superior handling of complex queries.

Table 1: Performance comparison on the QVHighlights validation split. The evaluation is segmented into ‘All’ for the entire validation set and ‘Long’ for queries exceeding 13 words.

Set	Method	MR		HD	
		R1 @0.5	@0.7	mAP Avg.	≥Very Good mAP
All	QD-DETR	62.84	46.77	41.23	39.49
All	Ours	67.61	50.65	44.80	40.98
Long	QD-DETR	58.44	39.94	38.40	38.84
Long	Ours	66.23	47.40	43.65	40.36

1 ROBUSTNESS TO QUERY LENGTH

Utilizing global text understanding, our method effectively predicts relevant frames even with longer text queries. To demonstrate its robustness, we evaluated it on queries exceeding 13 words, constituting 308 out of 1550 samples in the QVHighlights validation split. Results presented in Table 1 and Figure 1 show strong performance on extended queries. Notably, our model surpasses QD-DETR with a 4.77% improvement in R1@0.5 and a 3.57% increase in average mAP for the ‘All’ category. This margin expands to 7.79% in R1@0.5 and 5.25% in average mAP for the ‘Long’ category, highlighting our method’s robustness across varying query lengths.

2 ADDITIONAL ABLATION STUDIES

Number of layers. We explore various configurations of layer counts within our architecture to understand their impact on performance. The configuration with 2 cross-modal interaction layers, 3 encoder layers, and 3 decoder layers ($C = 2, E = 3, D = 3$) yields the best results.

Fine-Grained Alignment Loss. We explore the impact of varying the weights for fine-grained alignment loss, specifically the weights λ_{clip} and λ_{frame} . The results indicate that a balanced adjustment of

Table 2: Ablation results for different configurations of cross-modal interaction (C), transformer encoder (E), and decoder (D) layers on the QVHighlights validation set.

C	E	D	MR		HD	
			R1 @0.5	@0.7	mAP Avg.	≥Very Good mAP
2	2	2	66.45	49.55	43.28	40.65
2	3	3	67.61	50.65	44.80	40.98
3	2	2	67.03	49.10	43.41	40.87
3	3	3	64.58	48.71	43.70	40.33

Table 3: Ablation study results evaluating the impact of fine-grained alignment loss weights, λ_{clip} and λ_{frame} on the QVHighlights validation split.

λ_{clip}	λ_{frame}	MR		HD	
		R1 @0.5	@0.7	mAP Avg.	≥Very Good mAP
0.0	0.0	62.65	47.81	42.39	39.31
0.5	0.5	67.55	50.97	45.06	40.88
0.5	1.0	67.23	50.39	44.64	40.68
1.0	0.5	66.52	50.65	44.81	40.46
1.0	1.0	67.61	50.65	44.80	40.98

these weights does not significantly alter performance. Therefore, we opt for a weight configuration of $\lambda_{clip} = 1.0$ and $\lambda_{frame} = 1.0$, which consistently delivers optimal results.

3 FURTHER QUALITATIVE RESULTS

Figure 2 presents additional qualitative comparisons with our baseline, QD-DETR, highlighting the enhanced accuracy and context sensitivity of our approach.

117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174

175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232

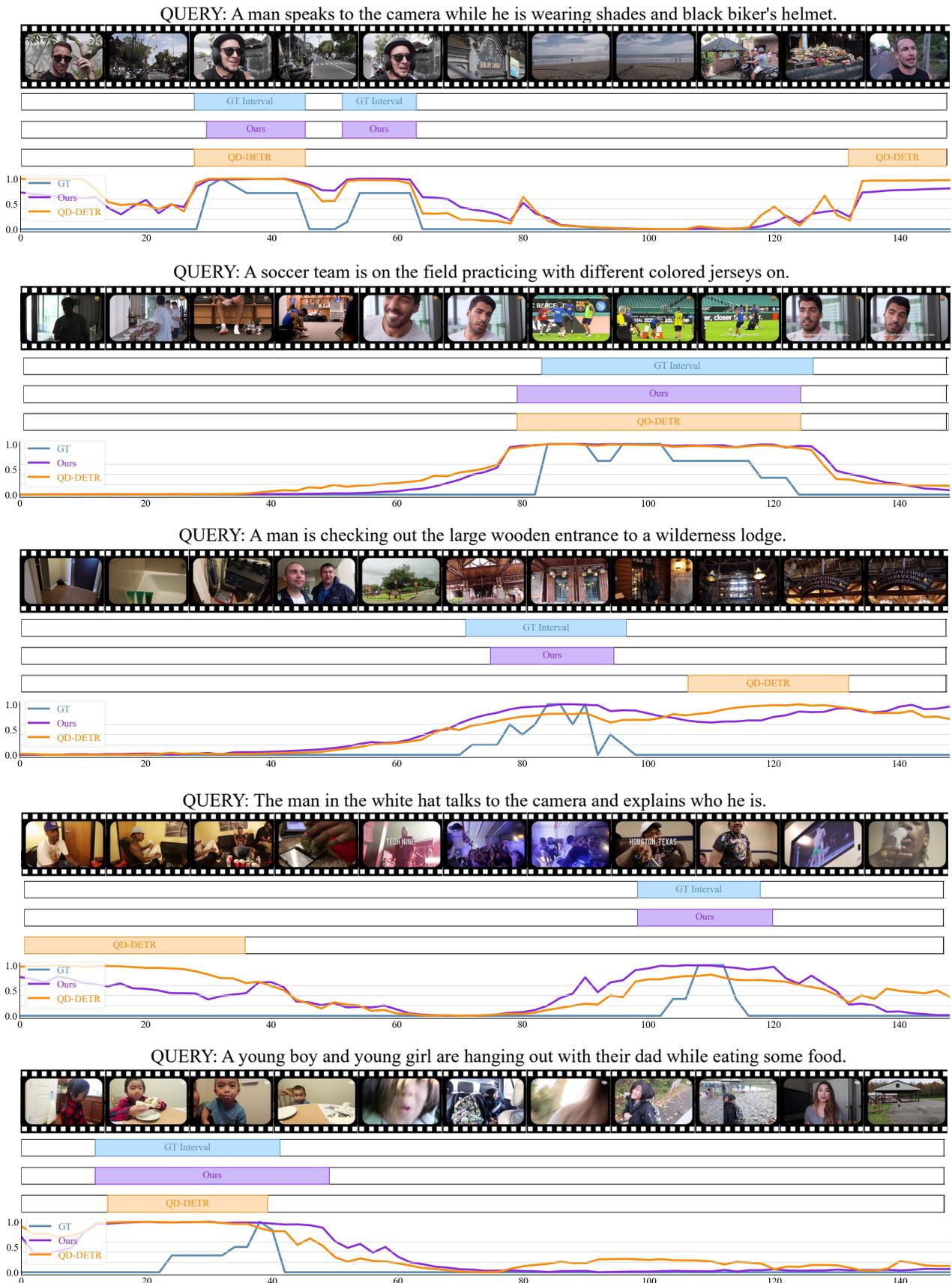


Figure 2: Extended qualitative results on the QVHighlights validation split, showcasing our method’s effectiveness in comparison to the baseline, QD-DETR. Displayed from top to bottom are the text queries, along with the corresponding predictions of moments and highlights for each method.