

Supplementary Materials: Prior Knowledge Integration via LLM Encoding and Pseudo Event Regulation for Video Moment Retrieval

Anonymous Author(s)

In this supplementary document, we offer additional examples to enhance comprehension of the method.

1 LLM ENDOERS AS RELATION REFINERS

We have validated the LLM encoders' ability to refine the inter-concept relations. This section presents more examples and illustrations to showcase their functionality on CLIP embeddings and in the VMR task.

1.1 How does it work on CLIP embeddings?

During the feasibility study, we examined how the LLM encoder refines inter-concept relations using triplets. To provide a clearer understanding, Fig. 1 presents more intuitive examples where we visualize the inter-conceptual similarity of CLIP embeddings before and after refinement using t-SNE. The figures illustrate that, after the refinement, the paired concepts are closer to each other (e.g., *car* and *road*), while the unpaired concepts are more distant (e.g., *car* and *printer*).

1.2 How does it work on VMR?

To demonstrate the process of relation refinement in the VMR task, we break down each query into individual concepts and utilize them as separate queries for VMR. We visualize the model's attention maps, showcasing its focus on both single-concept queries and the original queries (composed of combined concepts). From the Fig. 2 to Fig. 6, it is evident that with the LLM encoder, the model has demonstrated a better understanding on the composition of the concepts. The observations include:

Increased accuracy of subject identification. Upon completion of the refinement process, it becomes evident that the identification of individual concepts has significantly improved in terms of accuracy. This enhanced accuracy can be attributed to the increased consideration of context, as indicated by the refined inter-concept relations. Therefore, the predicted moments exhibit a comprehensive perspective that encompasses all concepts, rather than being dominated by visually prominent objects. For example, in Fig. 2, the post-refinement identification of the concept *Two guys* is markedly improved compared to a narrower moment range during the pre-refinement period. This is also observed in the identification of *competing* and *table tennis*. This observation indicates that the model possesses an improved comprehension of the scene, which is achieved through the collaborative efforts of the involved concepts. Another example in Fig. 3 demonstrates a refined precision in the detection of the concept *baby*. There is a misidentification of teenagers as *baby* when not applying refinement, but the proposed method is able to distinguish the similar semantics and localize the correct one. More examples can be found from Fig. 7, Fig. 4, and Fig. 5.

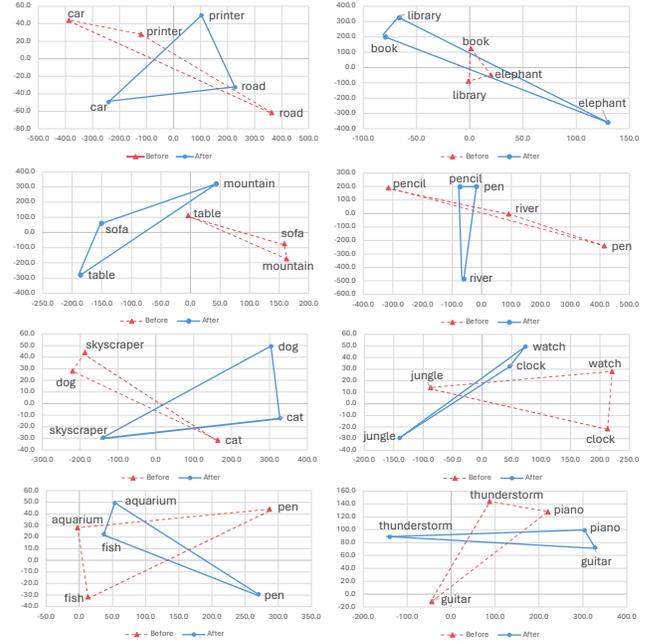


Figure 1: Visualization of the inter-conceptual similarity of the triplets before and after the refinement.

Elimination of the irrelevant dominance. Another observation is that the model gains awareness of background concepts such as scenes or events, allowing it to avoid biased focus solely on foreground concepts. In Fig. 6, the improved precision of the concept predictions is evident. In predictions made without the refiner, the results are influenced by the concept *car*, leading to an oversight prediction in a broader semantic context. However, with the refiner, the localization is narrowed, accurately encapsulating the *herd* as the focal point of the scene, thereby producing more precise predictions.

2 QUALITATIVE RESULTS

We provide qualitative results for QVHighlights in Fig. 8. Our method can more accurately locate moment boundaries without crossing them incorrectly, while also more accurately predicting the entire moment of the event, rather than just a portion of it.

ORIGINAL QUERY: *Two guys are competing at table tennis.*

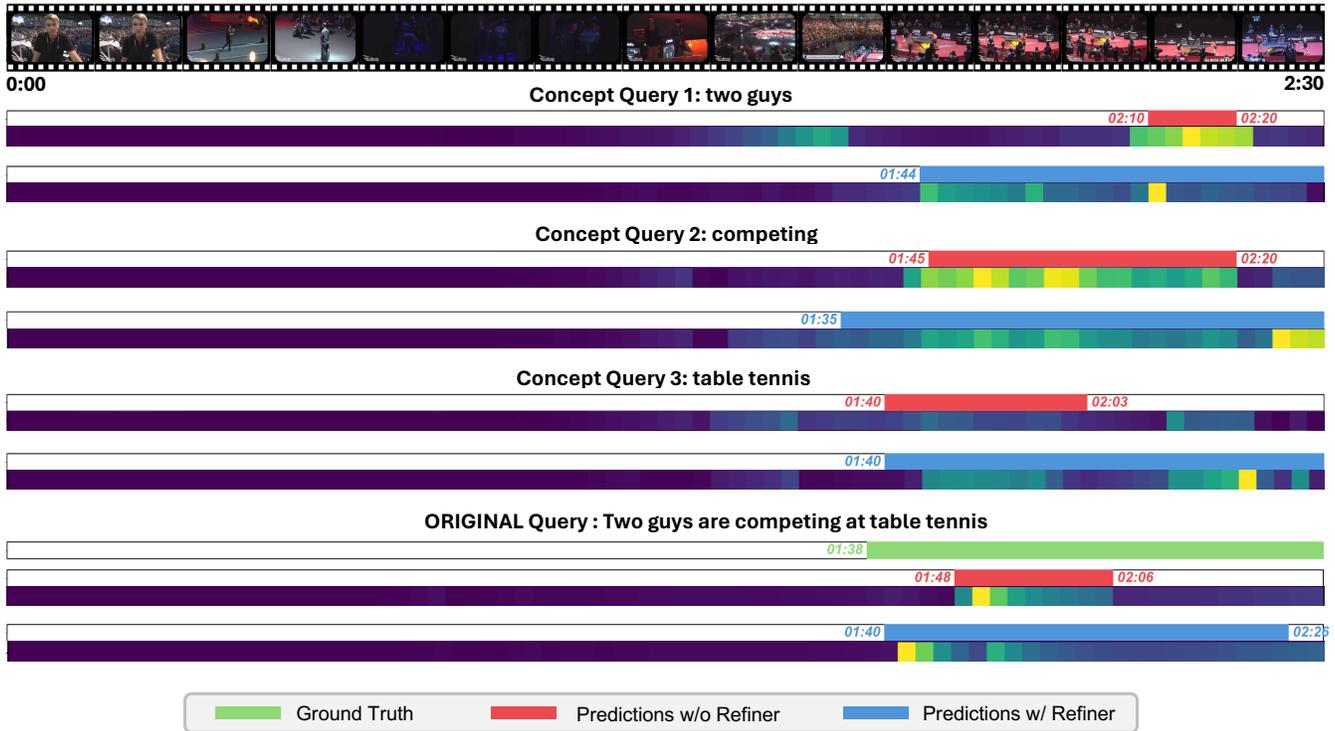


Figure 2: Visualization of Relation Refinement in the VMR Task. On the original query, the prediction made without the refiner is significantly affected by the dominant concepts *two guys* and *competing* (their prominence is evident from the high attention given to these concepts), while the contributions of other concepts are disregarded. By incorporating the refiner, the model has enhanced its comprehension of the collective semantics derived from multiple concepts. Notably, even subtle attention is given to each individual concept, with the refiner, the model has acknowledged the combined contributions of these concepts and ensured a more comprehensive coverage of the moment. This is an indication that the inter-concept relation has played a significant role during the inference.

233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290

291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348

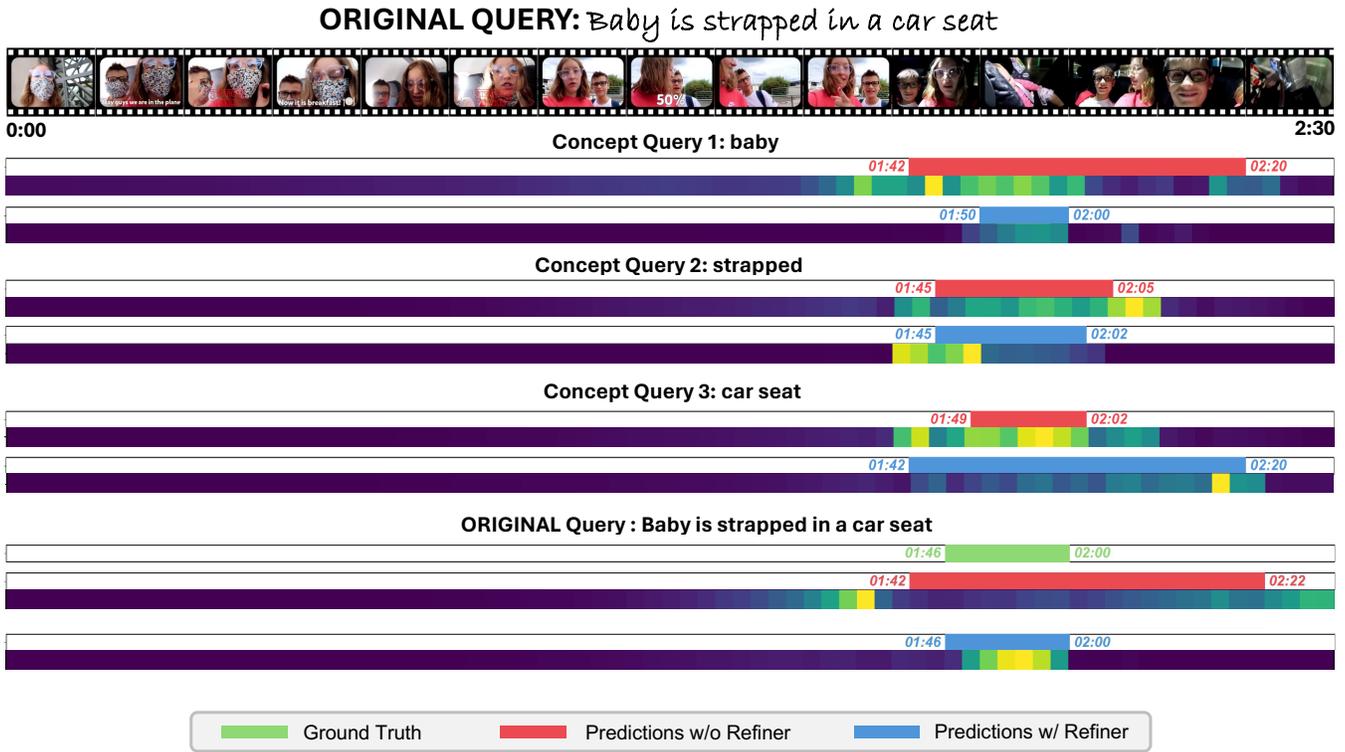
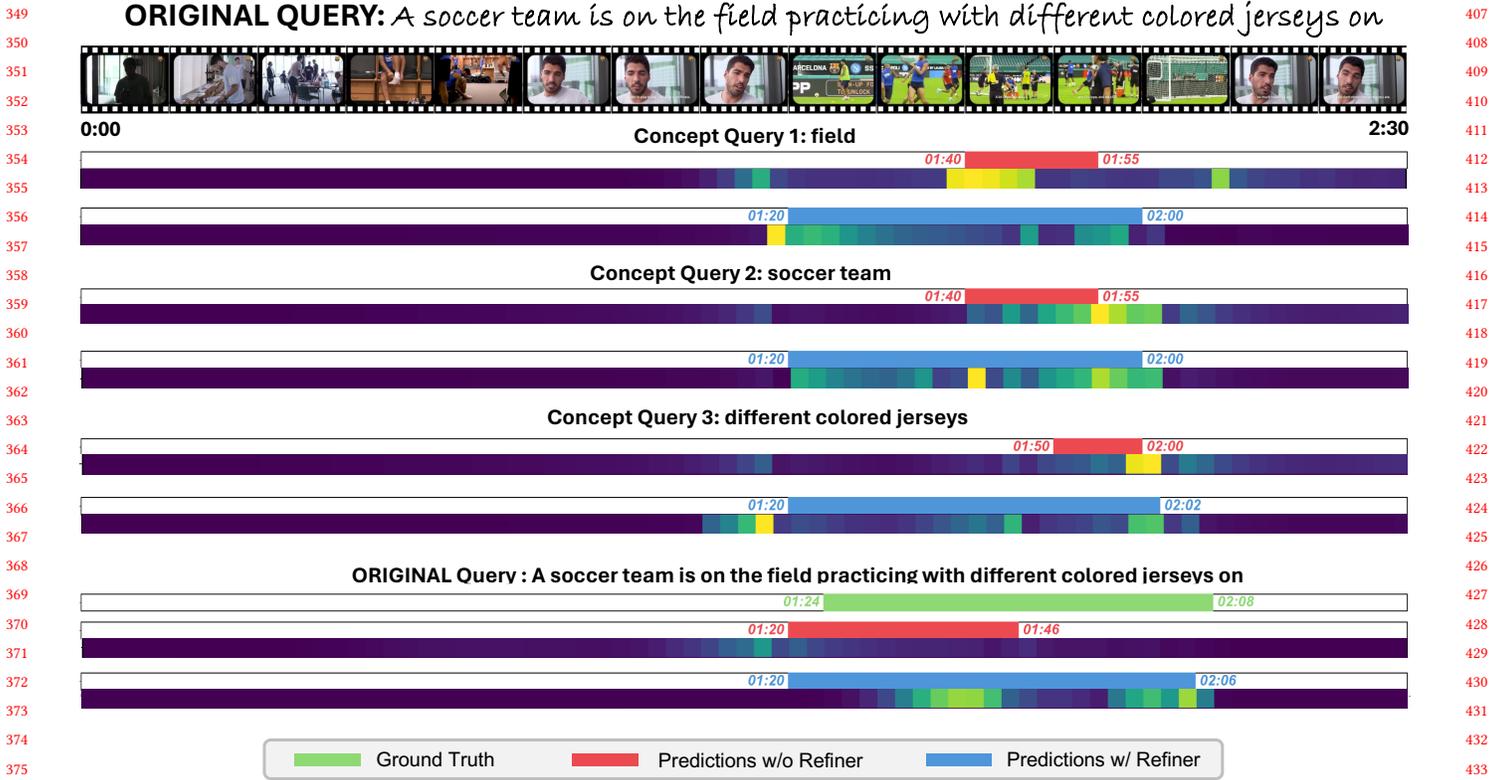
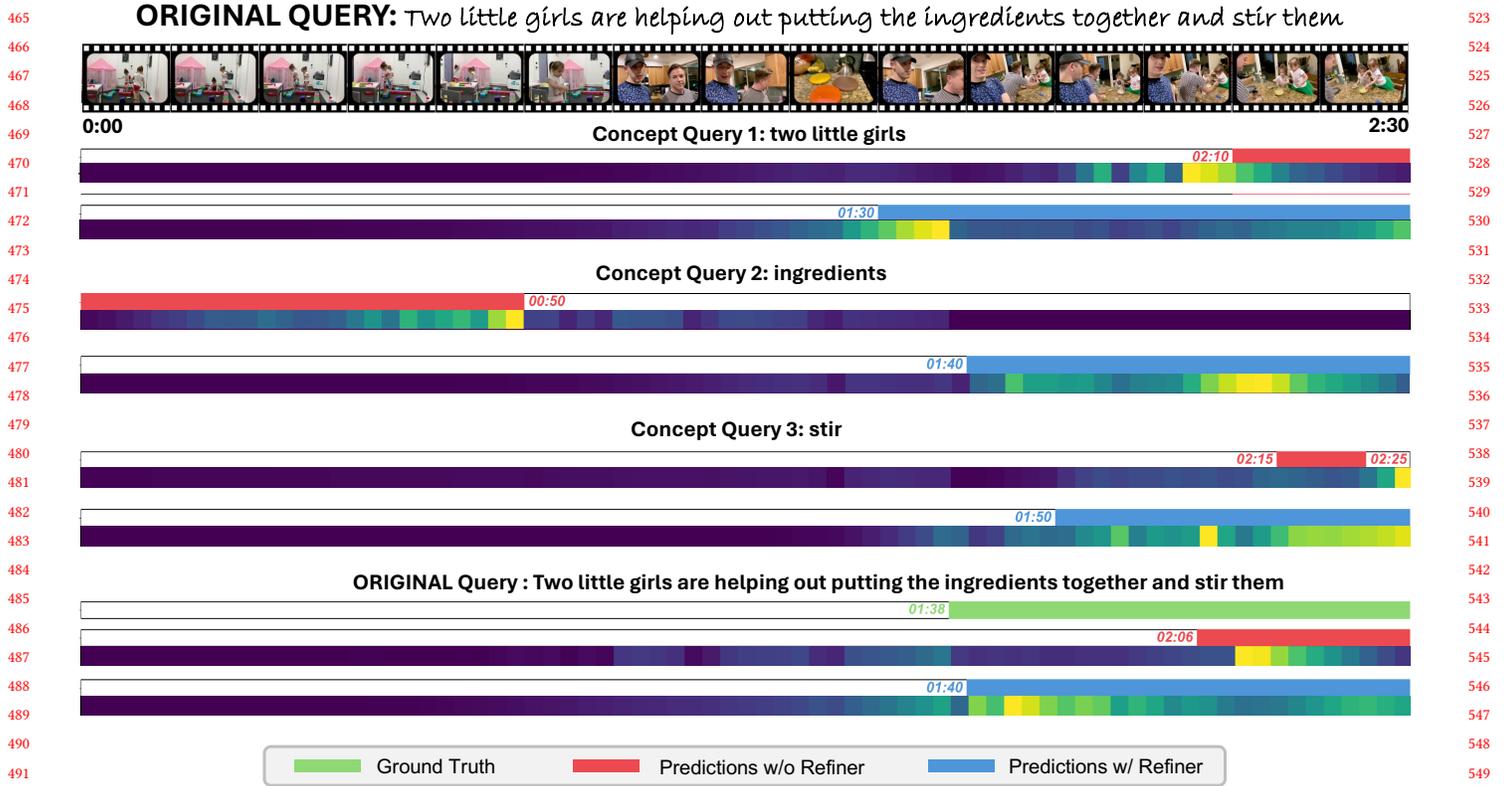


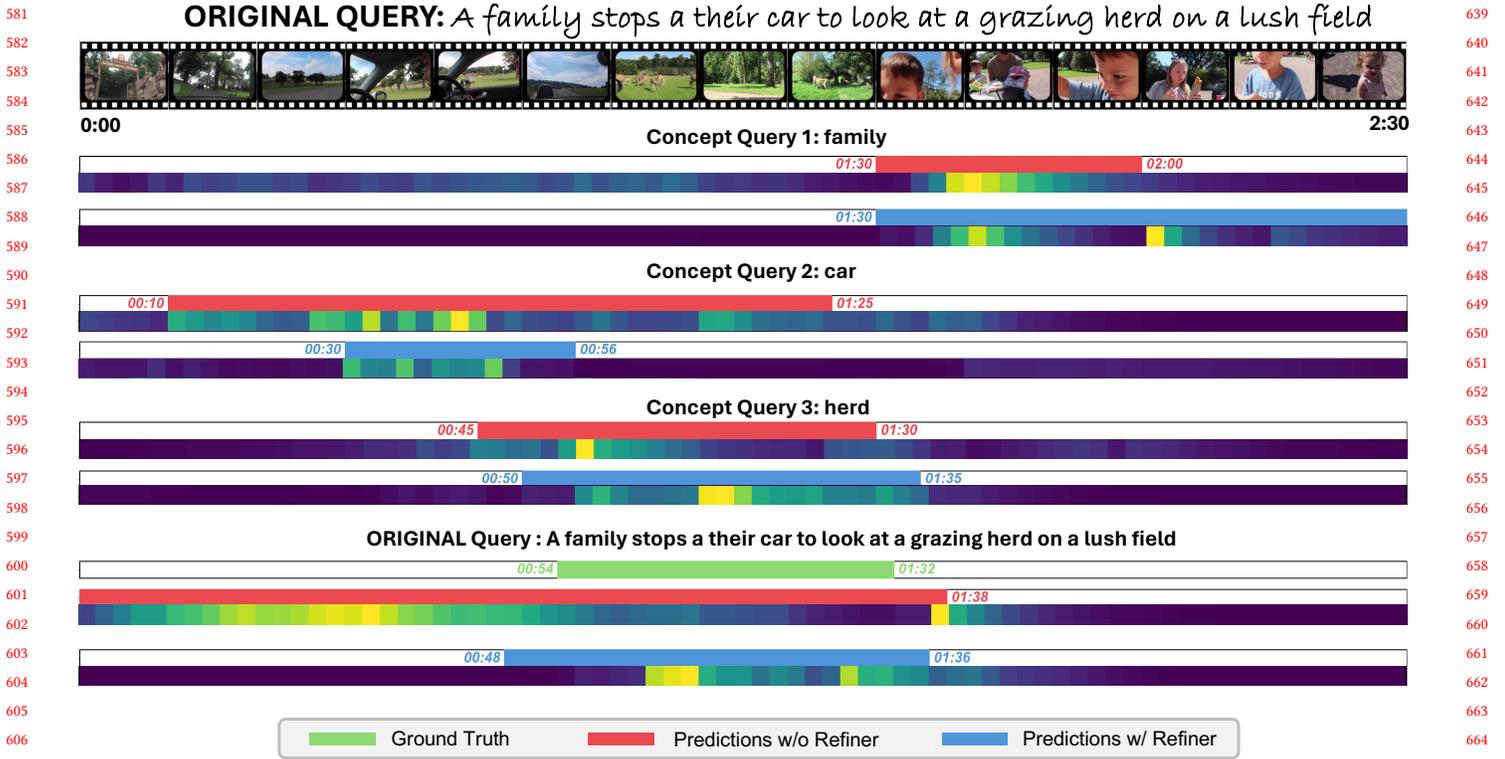
Figure 3: Visualization of Relation Refinement in the VMR Task. The primary concepts in this example are *baby* and *car seat*. Without the refiner, the model’s identification of the concept *baby* is distracted by the frequent appearance of teenagers, leading to a misleading moment alignment primarily focused on the segment containing the misidentified concept *baby*. However, with the refiner, the misidentification of concept *baby* is rectified. By contrast, the concept *car seat*, which was previously limited in scope due to its lack of visual prominence in the clip, exhibits a broader span after the refinement. This expansion suggests that contextual information (the in-car scene implied by other concepts) has been taken into account, reinforcing the significance of the concept *car seat*.



377 **Figure 4: Visualization of Relation Refinement in the VMR Task.** In the video, the concept *field* is the background, and the
 378 concept *soccer team* is the target objective of the video. The recognition accuracy of these two concepts controls the whole
 379 query corresponding to the ground truth. Without the refiner, the constant presence of the soccer team is interfered with the
 380 field in the background, resulting in both concepts being not fully recognized as individual concepts. This situation is solved
 381 with the refiner, which achieves a more detailed and comprehensive moment coverage.
 382
 383
 384
 385
 386
 387
 388
 389
 390
 391
 392
 393
 394
 395
 396
 397
 398
 399
 400
 401
 402
 403
 404
 405
 406



493 **Figure 5: Visualization of Relation Refinement in the VMR Task.** In this example, the action *stirring* is not visually prominent 551
 494 enough. This makes the model without the refiner neglect some of its presence. With the refiner, the model’s awareness of the 552
 495 context has been improved, resulting from the collective semantics provided concepts like *two girls*, *putting ingredients* and 553
 496 (potentially) other scene concepts not demonstrated. 554



608 **Figure 6: Visualization of Relation Refinement in the VMR Task.** In this example, the context is the concept *herd* and the main
 609 target is the concept *car* and *family*. Without the refiner, each concept is recognized accurately, but the composite prediction of
 610 the query is dominated by the concept *car*. With the refiner, attention is shifted to the composite context, which correctly
 611 prioritizes the *herd* in the scene and achieves a consistent fusion of concepts.
 612

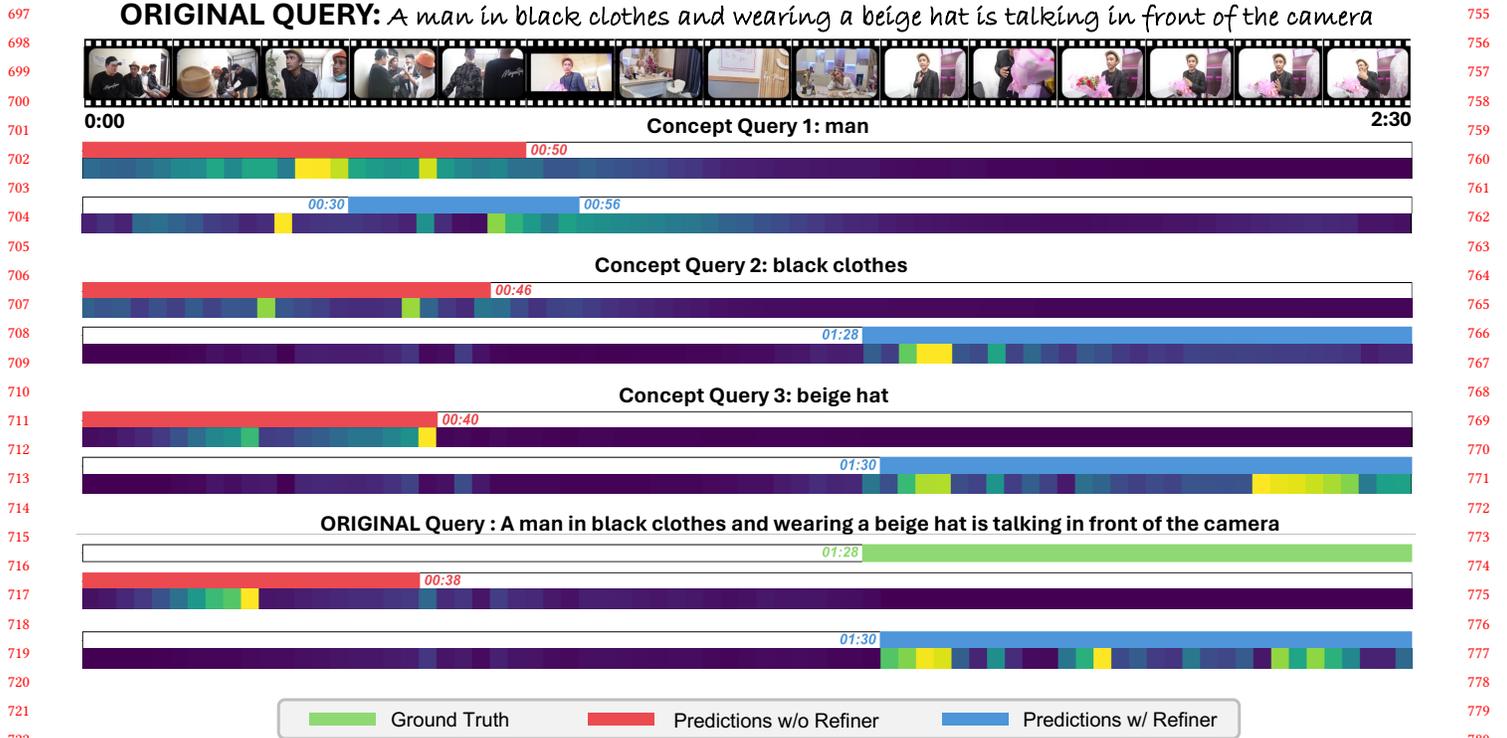


Figure 7: Visualization of Relation Refinement in the VMR Task. In this case, the most important concepts are *black clothes* and *beige hat*, while the concept *man* is a general target which appears throughout the video. Without the refiner, the concept *black clothes* and concept *beige hat* are both incorrectly identified, which leads to the final result being dominated by these two concepts. With the refiner, the identification of these two concepts is correct, which improves the overall accuracy of the query result. The model is more sensitive to *black clothes* and *beige hats*, thus does not conceptualize the *men* in the clip as the main query.

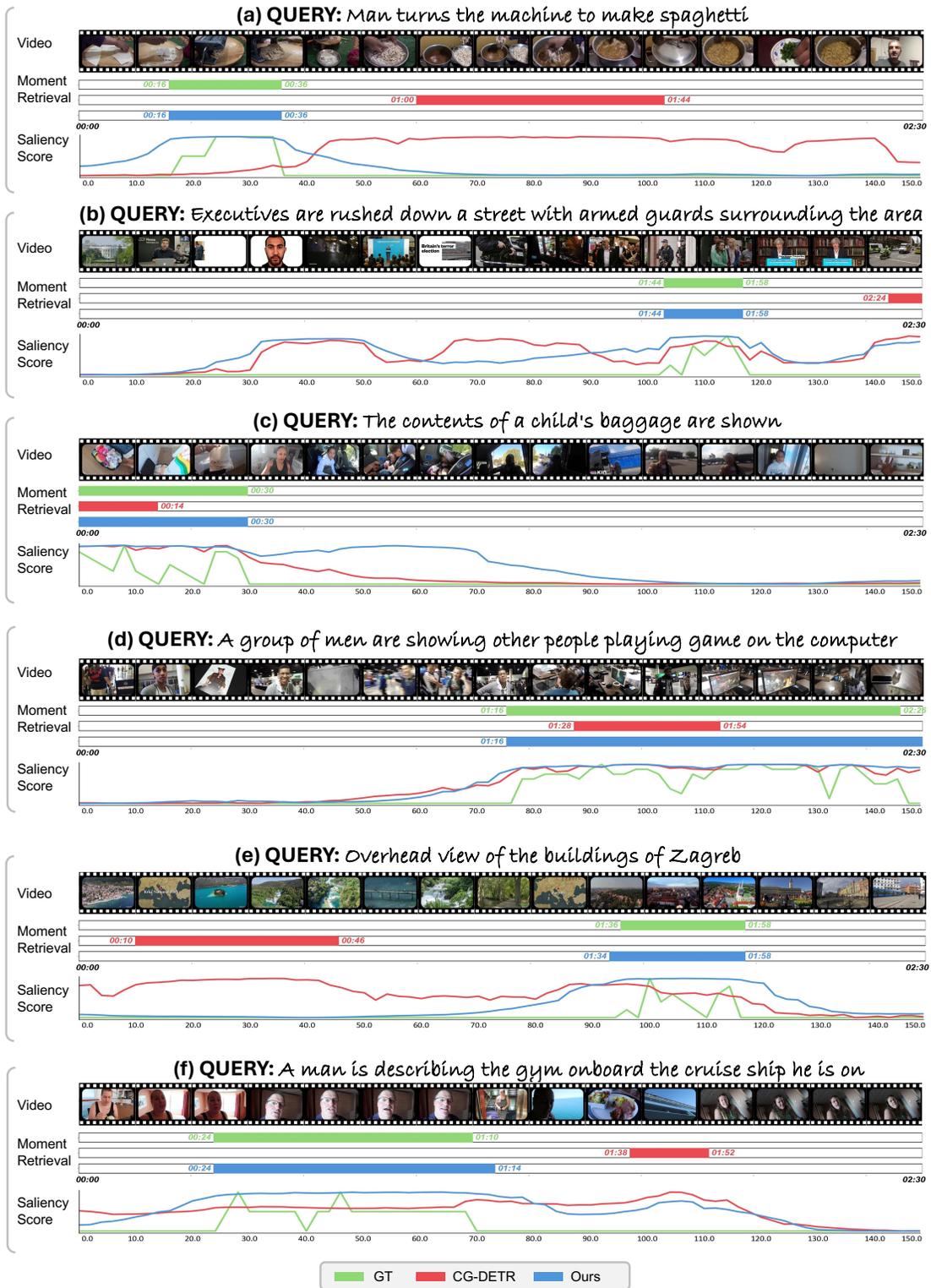


Figure 8: More visualizations of the joint moment retrieval and highlight detection results on QVHighlights val split. Our method accurately regresses moment boundaries and predicts highlight prominence scores with the help of the relation refiner.