

HieraMamba: Video Temporal Grounding via Hierarchical Anchor-Mamba Pooling

Supplementary Material

1. Additional Ablation Studies

We present additional ablation studies to isolate and assess the design choices of our proposed modules and losses. Unless otherwise specified, all experiments are conducted on Ego4D-NLQ using the base HieraMamba architecture (without auxiliary losses) to isolate component effects, with results averaged over five runs.

1.1. Anchor Generation Strategies

We evaluate four strategies for generating anchors within the AMP block. Given a temporal stride s , each anchor is computed from its corresponding s input tokens using one of the following pooling methods: (1) *Mean pooling*, which averages token features; (2) *Max pooling*, which selects the maximum activation per channel; (3) *Attention pooling*, which applies multi-head attention with a learnable query vector, following the attention pooling in CLIP [55]; and (4) *Gated pooling*, which adaptively blends mean- and max-pooled features via a learned gate.

Table 10 reports the performance of the base HieraMamba model when applying each pooling strategy to its AMP blocks. For a fair comparison, no additional ACC or SPC losses are applied, isolating the effect of the pooling strategy itself.

Interestingly, the best results are obtained with non-learned pooling methods (mean and max), with mean pooling slightly outperforming max pooling. In contrast, learned variants (attention and gated pooling) underperform, with attention pooling yielding marginally better results than gated pooling. This suggests that simple statistical aggregation produces more stable anchors by avoiding early information loss, allowing the AMP’s temporal modeling blocks (global and local encoders) to compress and extract the most salient content.

Pooling Method	R@1		R@5		Average R@1&5
	0.30	0.50	0.30	0.50	
Mean Pooling	18.23	12.55	39.13	28.78	24.68
Max Pooling	17.87	12.66	39.09	29.00	24.65
Attention Pooling	17.63	12.28	38.93	29.00	24.46
Gated Pooling	17.41	12.36	39.04	28.65	24.37

Table 10. Comparison of pooling methods on retrieval performance (R@1, R@5, and average of R@1 & R@5).

1.2. Impact of Pooling in Segment-Pooled Contrastive Loss

To assess the role of pooling in our Segment-Pooled Contrastive (SPC) loss, we compare the proposed pooled formulation (§4.4) with an *unpooled* variant. In the unpooled setup, rather than contrasting the pooled segment prototype $z_{\text{seg}}^{(l)}$ against all tokens in the ground-truth interval, we treat every in-segment token as an independent positive example. This removes the aggregation step, effectively forcing all tokens within the same ground-truth moment to be pulled tightly together in the embedding space.

Table 11 shows that the unpooled variant underperforms the pooled one, and even degrades the base model’s performance (HieraMamba without SPC or ACC losses). We attribute this drop to the fact that tokens within a ground-truth interval often correspond to distinct sub-actions (e.g., reaching, grasping, retracting) that should retain some temporal diversity. Forcing these heterogeneous sub-motions to collapse into a single point can blur fine-grained temporal dynamics, harming retrieval accuracy.

By contrast, our pooled formulation produces a holistic, high-level segment representation, which is then contrasted against positives and negatives at the segment level. This design preserves intra-moment variability while still providing strong query-level semantic guidance, encouraging ground-truth moments to be discriminative to surrounding, non-matching content.

Method	R@1		R@5		Average R@1&5
	0.30	0.50	0.30	0.50	
HieraMamba (base)	18.23	12.55	39.13	28.78	24.68
+ SPC Loss (Pooled)	18.52	13.01	39.99	29.39	25.23
+ SPC Loss (UnPooled)	17.23	11.77	38.95	28.24	24.05

Table 11. Comparison of SPC loss variants on retrieval performance.

1.3. Effect of ACC and SPC on Additional Datasets

We evaluate the contribution of the two contrastive losses across all datasets. Table 12 shows that each loss independently improves performance, and combining them achieves the strongest overall results.

2. Additional Implementation Details

We provide additional implementation details omitted from the main paper due to space constraints. Complete configu-

Dataset	Components		Recall (%) \uparrow				Avg.
	ACC	SPC	R1@0.30	R1@0.50	R5@0.30	R5@0.50	
Ego4D	\times	\times	18.23	12.55	39.13	28.78	24.68
	\checkmark	\times	18.52	13.24	39.62	29.50	25.22
	\times	\checkmark	18.52	13.01	39.99	29.39	25.23
	\checkmark	\checkmark	18.81	13.04	40.82	29.96	25.66
MADv2	\times	\times	14.36	8.70	27.38	18.73	17.29
	\checkmark	\times	14.82	9.19	28.17	19.43	17.90
	\times	\checkmark	14.86	8.63	27.32	18.60	17.35
	\checkmark	\checkmark	14.72	9.00	28.50	19.97	18.05
TACoS	\times	\times	58.29	49.16	83.10	72.08	65.66
	\checkmark	\times	58.84	48.26	83.90	73.88	66.22
	\times	\checkmark	58.99	48.64	82.78	73.21	65.81
	\checkmark	\checkmark	59.59	48.99	83.75	74.28	66.65

Table 12. Ablation of contrastive objectives across datasets.

rations and code are available in our official release.

2.1. AMP Details

As described in the main paper, we use Hydra [24] as the global encoder and a windowed Transformer [75] as the local encoder. For Hydra, we set $d_{\text{state}} = 64$, $d_{\text{conv}} = 7$, $\text{expand} = 2$, and $\text{head_dim} = 64$. For the local encoder, we configure a single layer ($\text{num_layers} = 1$) with a small attention window ($\text{window_size} = 5$), $n_{\text{heads}} = 2$, and $\text{stride} = 1$, enabling it to focus on very local context while remaining lightweight due to its minimal window size, head dimension, and depth. We stack these AMP blocks to construct the multi-scale video pyramid (Multi-Scale Video Encoder in Fig. 2, left), using 8, 8, and 9 layers for Ego4D, TACoS, and MAD, respectively.

2.2. Training Details

We adopt the same training and inference settings as prior work [49], including learning rate, number of epochs, and other hyperparameters. Below, we detail the moment decoding procedure and the loss functions used for optimization.

Moment decoding. At each scale l , the refined sequence $\tilde{V}^{(l)} = \{\tilde{\mathbf{v}}_t^{(l)}\}_{t=1}^{L_l}$ is passed through two lightweight heads (three 1D convolutions each): (i) a classification head that outputs a confidence score $p_t^{(l)}$, and (ii) a regression head that predicts normalized start/end offsets $\delta_t^{(l)} = (\delta^s, \delta^e)$. For brevity, we omit (t, l) when clear from context.

Given the effective stride $S^{(l)}$ (e.g., $S^{(l)} = s^{l-1}$ for geometric downsampling by s), each token produces a proposal

$$\hat{\mathbf{y}} = (S^{(l)}(t - \delta^s), S^{(l)}(t + \delta^e)).$$

We rank all proposals across t and l by p , and apply Soft-NMS [3] over the multi-scale set to merge overlapping candidates, following common practice in video grounding [49, 75]. The final output consists of the top- k moment

predictions $\{(t_s, t_e)\}_{k=1}^K$ after Soft-NMS re-ranking.

Training objectives. The model is optimized with three loss terms: (i) a classification loss \mathcal{L}_{cls} using Focal Loss [39], (ii) a regression loss \mathcal{L}_{reg} using Distant IoU Loss [79], and (iii) a contrastive loss $\mathcal{L}_{\text{contrast}}$ that combines the proposed ACC and SPC losses. $\mathcal{L}_{\text{contrast}}$ is as defined in Eq. 8 of the main paper, which are controlled by λ_{ACC} and λ_{SPC} . We set $(\lambda_{\text{ACC}}, \lambda_{\text{SPC}})$ to (10, 1) for Ego4D, (1, 0.1) for TACoS, and (0.5, 0.6) for MAD. The final training objective is

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{contrast}}.$$

3. Qualitative Results.

In this section, we present qualitative visualizations of our model’s predictions for diverse language queries across a variety of scenarios. We use the Ego4D-NLQ [14] benchmark, where the ground-truth moment length can range from as short as one second to over 30 seconds, depending on the query and scenario. We first compare our visualizations against those from SnAG [49], the state-of-the-art open-source model for which we can run experiments, then provide additional visualizations showcasing our own predicted moments.

3.1. Qualitative Comparison with State-of-the-Art

Figure 5 presents a side-by-side qualitative comparison between SnAG [49] and our HieraMamba model. Each colored bar corresponds to a different language query for a given video clip: the yellow segment marks the ground-truth moment, the blue segment (beneath the yellow) shows SnAG’s prediction, and the green segment (final row) depicts our prediction.

The examples span diverse scenarios from the Ego4D-NLQ benchmark, where ground-truth moments range from fleeting events lasting barely a second to extended activities exceeding 30 seconds. This diversity demands a model capable of reasoning over both fine-grained and long-range temporal contexts. By leveraging hierarchical semantic representations across multiple temporal scales, our model effectively adapts to this variability—capturing the precise span for short events while maintaining coherence for extended activities.

In many cases, SnAG’s predictions exhibit partial misalignment with the ground truth, starting too early, ending prematurely, or drifting away from the relevant content. In contrast, HieraMamba’s predictions remain closely aligned with the annotated intervals across all temporal ranges. For example, in Query 2 of the first clip, SnAG localizes the moment too early, omitting critical visual evidence, whereas our method covers the complete span. Similarly, in the clothing store example, our prediction preserves the full interaction interval, avoiding the truncation seen in SnAG’s

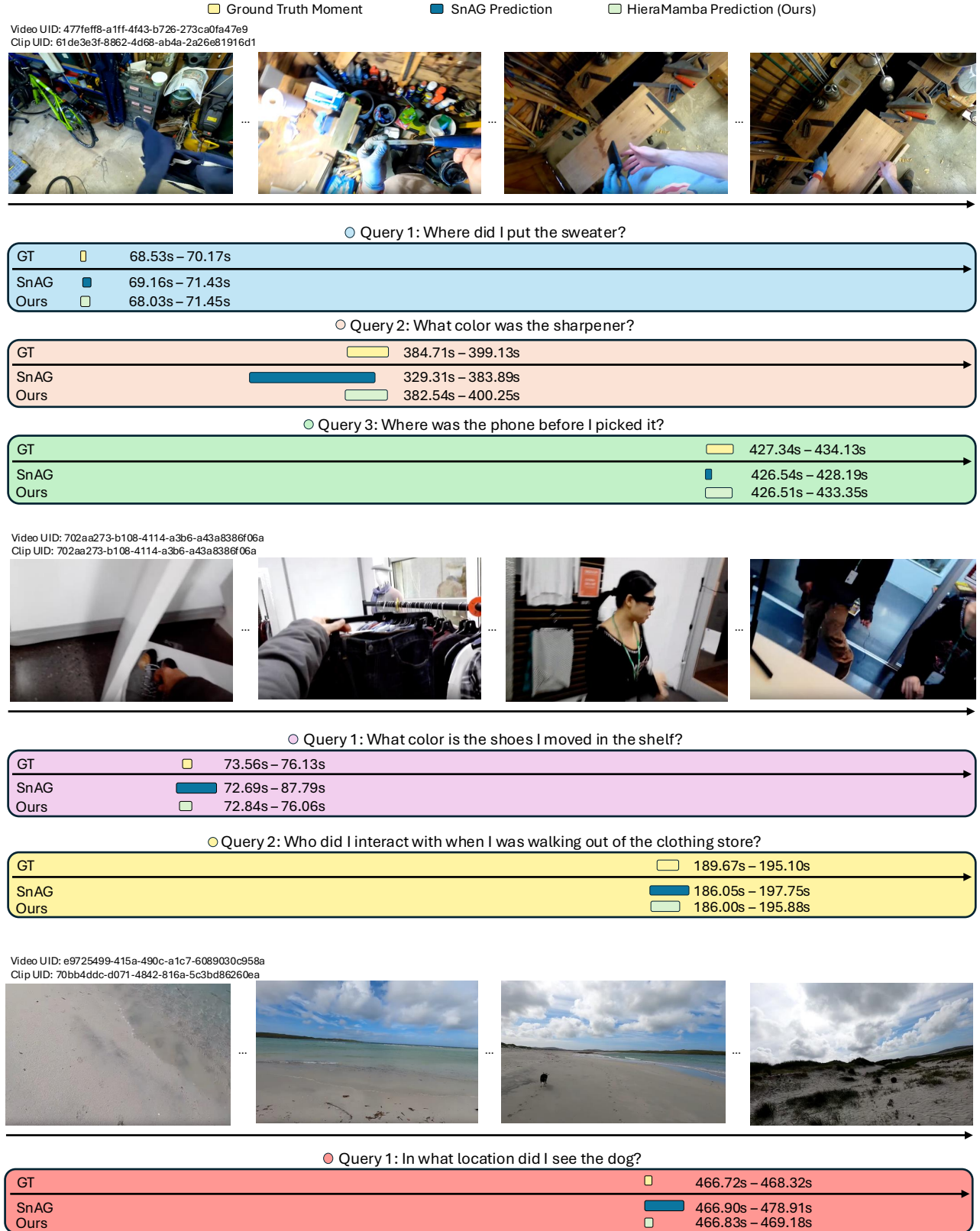


Figure 5. Qualitative Results Comparison with SnAG [49].

output. Even in cases where both predictions are close to the ground truth (e.g., second query in the clothing store scenario), our boundaries are slightly more precise, reflecting improved temporal alignment.

Overall, these qualitative results illustrate how multi-scale temporal reasoning enables HieraMamba to robustly localize events of vastly different durations, providing faithful and semantically coherent grounding across a wide variety of queries and scenarios.

3.2. Qualitative Results: Handling Diverse Temporal Granularities

Figures 6 and 8 show qualitative examples from Ego4D-NLQ demonstrating our model’s ability to localize moments of vastly different durations, even within the same continuous video. In realistic egocentric recordings, multiple queries can refer to events at very different temporal scales: a brief action lasting about a second (e.g., picking up an item) may appear alongside an extended activity exceeding 30 seconds (e.g., a multi-step cooking or interaction sequence). This variation arises not only across different videos, but also frequently within the same video, making accurate localization particularly challenging.

HieraMamba addresses this challenge by producing semantically rich representations at multiple temporal scales—capturing fine-grained details for short moments while also maintaining coherent long-range context for extended activities. This multi-scale representation enables the model to adapt its grounding behavior based on the temporal demands of each query, without sacrificing precision for short events or coverage for long events.

As shown in the figures, our predictions align closely with the ground truth across a wide range of temporal granularities. For short-duration queries, boundaries are tightly matched to the relevant frames; for long-duration queries, the predicted segments span the full relevant context without truncation or drift. These highlight our model’s ability to seamlessly navigate between fine and coarse temporal reasoning, a capability essential for handling the mixed temporal demands present in real-world scenarios.

3.3. Qualitative Results: Failure Cases

Figures 7 show qualitative failure examples from Ego4D-NLQ [14]. For the first query, “What color is the hammer?”, both intervals contain the same hammer, but the predicted moment corresponds to the hammer being actively used rather than stationary. The model appears biased toward dynamic usage cues instead of static appearances.

For the second query, “What tool did I use on the engine?”, the model predicts a much longer interval because the tool and engine remain in view well beyond the actual interaction. This suggests that the model conflates object visibility with tool usage.

For “What color was the bucket next to the door?”, the ground-truth bucket is upside-down and partially visible at the frame edge, requiring implicit 3D reasoning. The predicted segment instead shows another bucket in a clearer, more canonical view, which the model favors over the subtler ground-truth configuration.

For the interpersonal query “Who was with me when I repaired the engine?”, the model truncates the interval when the other person becomes temporarily occluded due to camera motion, revealing a lack of persistent 3D environment understanding.

For “Where was the circular metal before I picked it up?”, the model predicts a moment involving a small idler pulley, likely confusing it with the referenced circular metal due to similar shape and cluttered tool context.

Finally, for “Where was the screwdriver before I picked it up?”, the model focuses on a moment where a screwdriver is visible near a dead-blow hammer being lifted. The similar shape and nearby motion lead the model to incorrectly associate the action with the screwdriver.

These failure cases highlight several systematic limitations. The model often conflates visibility with action boundaries, struggles with fine-grained object discrimination, and treats temporarily occluded objects or people as absent. It also favors dynamic tool usage over static attributes and has difficulty interpreting non-canonical viewpoints that require implicit 3D reasoning. Together, these patterns suggest that achieving reliable moment localization will require better modeling of action semantics, object identity, occlusion, and spatial geometry.

4. Limitations

While HieraMamba provides a scalable and accurate framework for long-video temporal grounding, it also has several limitations that open avenues for future work. First, although our model achieves linear-time complexity and supports multi-scale reasoning, it relies on frozen video backbones. This modular design offers flexibility in selecting video encoders but also decouples video feature learning from the temporal grounding objective. Jointly fine-tuning the video backbone together with our model could further improve performance, though at the expense of the substantial compute required for training large backbone models.

Second, our anchor generation strategy operates with a fixed temporal stride. An adaptive mechanism that adjusts the stride dynamically based on video content, allocating more anchors to regions with higher temporal density and fewer to less informative segments, could further enhance localization accuracy and efficiency.

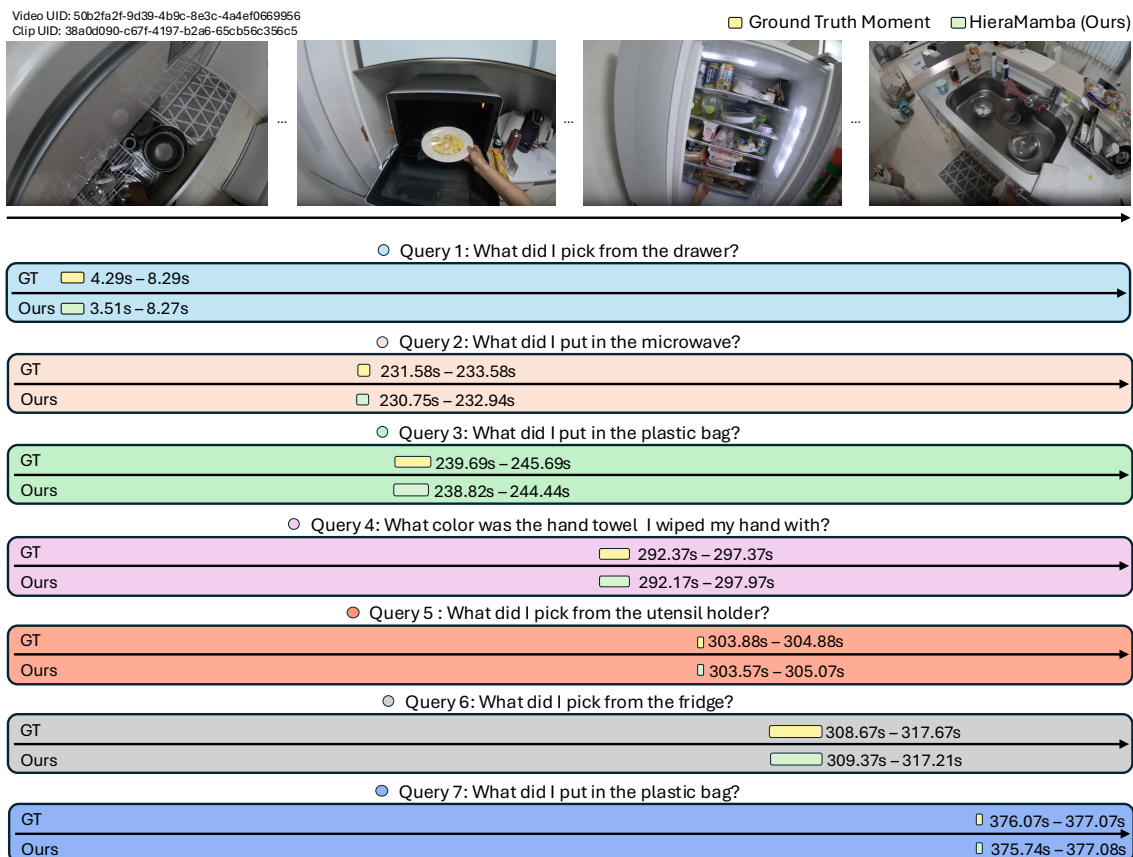


Figure 6. Qualitative Results

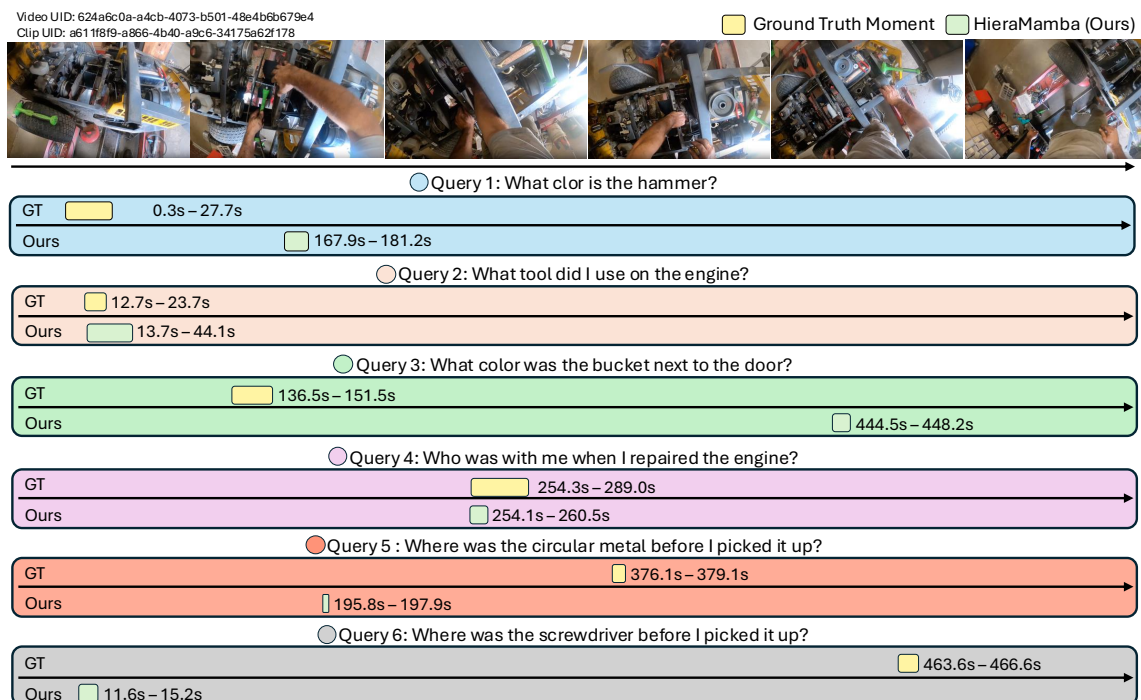


Figure 7. Qualitative Failure Examples.

Video UID: 9f28e782-417c-4c8b-a7ae-42fc96a0e94f
Clip UID: d7b8f461-db42-4365-9f89-83f923528293

Ground Truth Moment HieraMamba (Ours)



Query 1: Where was the chopping board before I dropped it?

GT 0.67s – 5.43s

Ours 0.00s – 5.39s

Query 2: How much much oil did I put in the pan?

GT 50.96s – 53.85s

Ours 51.09s – 54.40s

Query 3: What did I put in the fridge?

GT 107.03s – 111.03s

Ours 106.87s – 111.20s

Query 4: How much meat did I put in the pot?

GT 256.42s – 259.78s

Ours 257.00s – 260.34s

Query 5: How much soap did I apply on the sponge?

GT 285.84s – 287.33s

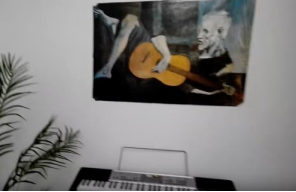
Ours 285.78s – 287.76s

Query 6: How many tomatoes did I chop?

GT 446.46s – 467.34s

Ours 443.42s – 469.49s

Video UID: 262fd2ec-8eba-44d9-8082-4d0574f7a515
Clip UID: 38a0d090-c67f-4197-b2a6-65cb56c356c5



Query 1: In what location did I see a wall decoration?

GT 3.73s – 35.93s

Ours 0.00s – 37.71s

Query 2: Where was a flower before I smelled it?

GT 172.28s – 179.42s

Ours 172.00s – 178.87s

Query 3: How many cans were in the fridge?

GT 315.84s – 326.02s

Ours 313.58s – 326.86s

Figure 8. More qualitative results.