Adaptive Query-Conditional Fusion for Egocentric Natural Language Query Grounding

Enmin Zhong, Carlos R. del-Blanco, Fernando Jaureguizar, Narciso García

Grupo de Tratamiento de Imágenes (GTI), Information Processing and Telecommunications Center ,
ETSI Telecomunicación, Universidad Politécnica de Madrid, Spain
{enmin.zhong, carlosrob.delblanco, fernando.jaureguizar, narciso.garcia}@upm.es

Abstract. Egocentric Natural Language Query (NLQ) grounding aims to localize, within a long first-person video, the temporal interval that answers a free-form text query. Not all queries rely on the same visual evidence: a manipulation query is answered by hand kinematics, a social query by appearance, and a localization query by scene context. Existing systems, however, fuse auxiliary modalities uniformly across all query types and ignore the semantic structure that the queries themselves encode. We argue that fusion should instead be conditioned on the query, so that each evidence source contributes only when it is relevant. Building on this principle, we introduce hand trajectory as an auxiliary modality that complements video and text for NLQ grounding, particularly informative for manipulation-centric queries, and a Query-Conditional Fusion module that weights it according to query content. On Ego4D NLQ, our method attains the best overall performance and improves or matches uniform-fusion methods across all query categories.

Keywords: Egocentric video · Natural language query grounding · Hand-Object Interaction · Temporal grounding · Multimodal fusion · Ego4D

1 Introduction

Wearable cameras are turning first-person video into a searchable record of daily life. Being able to retrieve, from hours of egocentric footage, the precise moment when a specific event occurred—"where did I put my keys?", "who handed me the document?"—by asking in natural language would unlock applications across assistive technology, industrial training, and everyday memory augmentation. Egocentric Natural Language Query (NLQ) grounding, formalised on Ego4D [4], operationalises this goal: given a long first-person video and a free-form text query, a model must predict the temporal span $[t_s, t_e]$ that answers it. Strong baselines such as GroundNLQ [5] address egocentric NLQ grounding by combining three main components: pretrained video features, text-query features, and a temporal localization head. In particular, they extract visual representations from large pretrained video encoders such as InternVideo [12] and EgoVLP [6],

encode the natural-language query using CLIP [10] text features, and then feed the resulting video–text representation into a transformer-based prediction head to estimate the temporal segment that answers the query. This combination of strong video-language representations and transformer-based temporal reasoning has produced competitive results on Ego4D NLQ. UniVTG [7] and R2-Tuning [9] further improve temporal language grounding from complementary perspectives. UniVTG proposes a unified video-language grounding framework that benefits from large-scale pretraining and can address multiple temporal video understanding tasks, including moment retrieval, highlight detection, and video summarization. In contrast, R2-Tuning focuses on efficient adaptation: instead of fully fine-tuning large video-language models, it selectively adapts temporal components to transfer image-pretrained representations to video grounding more effectively.

Subsequent works have extended GroundNLQ with additional modalities. GazeNLQ [8] adds predicted gaze signals to highlight where the camera wearer is likely attending. ObjectNLQ [2] introduces object detections to make the model more aware of relevant entities in the scene. OSGNet [1] further enriches the representation by adding both object-level information and shot-level branches, improving the model's ability to localize query-relevant moments in egocentric video.

In contrast to gaze, object, and shot-level cues, hand trajectory has received comparatively little attention as an explicit auxiliary modality for NLQ grounding, even though a substantial portion of Ego4D NLQ queries are closely tied to hand–object manipulation. Queries involving hand–object interaction, object state changes, quantities, or object locations often require identifying moments in which the camera wearer reaches for, grasps, moves, places, or releases an object. Hand priors are already well established in adjacent egocentric tasks, such as hand–object contact detection [11] and action anticipation [3]. However, to the best of our knowledge, hand skeleton trajectories has not been explicitly exploited as an auxiliary signal for the NLQ grounding task.

A first contribution of this work is the use of hand kinematics as an additional source of information to improve temporal grounding prediction. Specifically, sequences of hand skeletons are encoded to capture the spatial and temporal relationships among hand-joint trajectories. This encoding is designed to be aligned with the video frame structure, so that the resulting hand-trajectory features can be naturally fused with the visual video representation at later stages of the model.

Existing auxiliary-modality systems usually fuse all information sources in the same way for every query. However, the query itself indicates which evidence is most relevant: hand trajectories for manipulation, appearance for people-related queries, or scene context for object-location queries. Therefore, fusion should be query-dependent rather than uniform.

The second contribution of this work is a Query-Conditional Fusion mechanism that determines how the available information sources should be combined for each query. The proposed approach jointly leverages three sources of infor-

mation — video appearance, textual query features, and hand-trajectory cues — by first pairing the video with each auxiliary modality to obtain two bimodal representations, and then adaptively combining these two representations according to the query. The selection is performed by a weighting mechanism that takes the query embedding as input and assigns higher weight to the bimodal representation that is most appropriate for the query content. For example, manipulation-related queries can be routed toward hand-trajectory information, whereas queries focused on people, objects, or scene context may rely more on visual or textual representations. Overall, this query-adaptive fusion strategy allows the model to exploit each modality only when it is useful, improving temporal grounding while avoiding the noise introduced by irrelevant auxiliary information.

2 Method

The proposed system takes an egocentric video clip and a natural-language query as input and predicts the temporal answer span $[t_s, t_e]$. As illustrated in Fig. 1, the architecture consists of four modules: *Unimodal Encoders*, *Bimodal Encoders*, *Query-Conditional Fusion*, and *Temporal Segment Prediction*. The *Unimodal Encoders* project the input video, the query text, and the camera-wearer's hand trajectory into modality-specific embeddings, denoted by E_v, E_t, and E_h, respectively. The visual embedding E_v is then combined in parallel with the textual embedding E_t and the hand-trajectory embedding E_h by the *Bimodal Encoders*, producing the visual-semantic representation E_{vt} and the hand-motion-aware representation E_{vh}. The visual-semantic representation E_{vt} incorporates query-relevant textual information into the video representation, allowing the model to emphasize visual evidence that is semantically aligned with the natural-language query. In contrast, the hand-motion-aware representation E_{vh} enriches the video representation with hand-trajectory cues, which are particularly useful for queries involving object manipulation or hand–object interaction. The *Query-Conditional Fusion* module then adaptively combines E_{vt} and E_{vh} according to the content of the query. This design is motivated by the fact that hand-trajectory information is highly informative for some queries but irrelevant, or even misleading, for others. Therefore, the fusion mechanism adaptively weights each bimodal representation depending on its expected usefulness for the current query. The resulting adaptively fused video embedding E_o is finally processed by the *Temporal Segment Prediction* module, which uses a multi-scale temporal pyramid approach to predict accurately the start and end boundaries of the answer span.

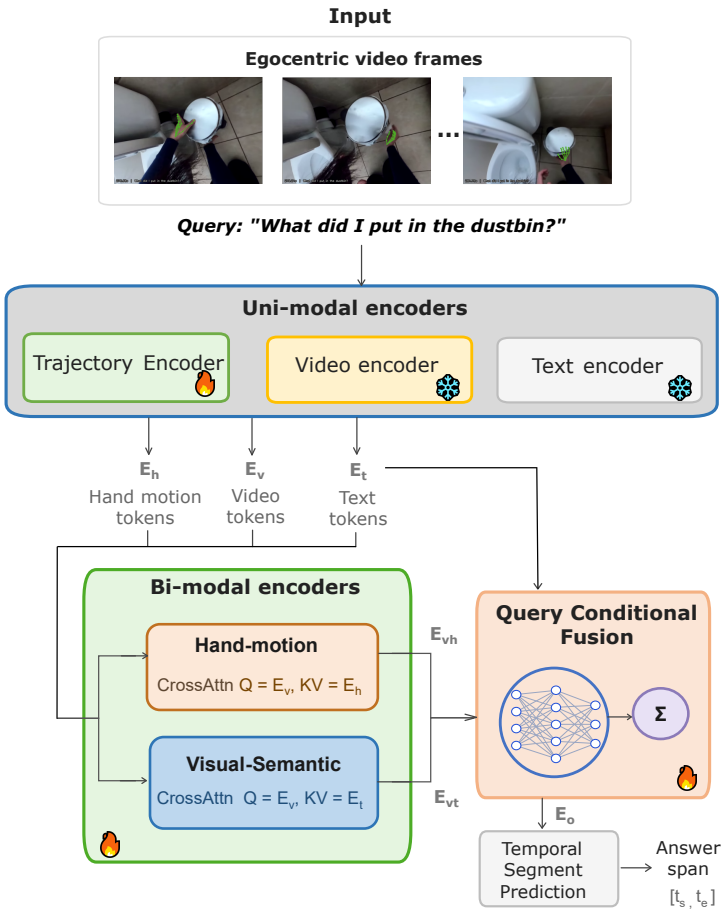

Fig. 1. Overall architecture. Given an egocentric clip and a natural-language query, the *Unimodal Encoders* produce video tokens $\mathbf{E_v}$ (frozen InternVideo [12] and EgoVLP [6]), text tokens $\mathbf{E_t}$ (frozen CLIP [10]), and hand-motion tokens $\mathbf{E_h}$ (trainable Trajectory Encoder). The *Bimodal Encoders* fuse the video with each auxiliary stream via cross-attention, yielding the visual-semantic ($\mathbf{E_{vt}}$) and hand-motion-aware ($\mathbf{E_{vh}}$) representations. *Query-Conditional Fusion* combines both into $\mathbf{E_o}$ with query-dependent weights, from which *Temporal Segment Prediction* localizes the answer span $[t_s, t_e]$.

2.1 Unimodal Encoders

To extract complementary information from the different input sources, we first project the video, the natural-language query, and the hand trajectory into a shared embedding space using dedicated encoders.

The module produces modality-specific embeddings of dimension D, denoted by $\mathbf{E}_v$, $\mathbf{E}_t$, and $\mathbf{E}_h$, respectively. The Video Encoder processes the egocentric video clip and produces a sequence of T_v visual tokens $\mathbf{E}_v \in \mathbb{R}^{D \times T_v}$ using

pretrained and frozen InternVideo [12] and EgoVLP [6] backbones. The Text Encoder maps the natural-language query into a sequence of T_t textual tokens $\mathbf{E}_t \in \mathbb{R}^{D \times T_t}$ through a pretrained and frozen CLIP [10] text encoder.

Finally, a trainable Trajectory Encoder processes temporal sequences of hand skeletons extracted from the raw video using a hand skeleton extractor [13], generating video-aligned kinematic representations $\mathbf{E}_h \in \mathbb{R}^{D \times T_v}$. More specifically, the hand skeleton extractor generates for each frame 21 three-dimensional joint coordinates for each hand. To ensure temporal alignment with the visual tokens $\mathbf{E}_v$, the skeleton sequence is divided into non-overlapping windows of 16 frames; each window is encoded through a lightweight 3D-CNN that captures local spatio-temporal joint interactions, followed by a Transformer encoder that models long-range temporal dependencies across the trajectory [14].

2.2 Bimodal Encoders

This module combines the visual representation with two complementary sources of information: the natural-language query and the hand-trajectory representation (see Fig. 1). Its goal is to enrich the video tokens with the information needed to temporally localize the segment that answers the query. The text query specifies *what* the model should look for, whereas the hand trajectory provides additional motion cues that can be especially informative for queries involving hand–object interactions.

To integrate these sources, we use two Transformer-based *bimodal encoders*. The first one, the *video-text encoder*, injects query semantics into the visual representation producing the token embeddings $\mathbf{E}_{vt}$. It uses the visual tokens $\mathbf{E}_v$ as queries Q, and the text tokens $\mathbf{E}_t$ as keys K and values V in a cross-attention module. This allows each video token to attend to the parts of the query that are most relevant for interpreting the visual content.

The second one, the *video-hand encoder*, incorporates hand-kinematic information into the visual representation yielding the token embeddings $\mathbf{E}_{vh}$. In this case, the visual tokens $\mathbf{E}_v$ are again used as queries Q, while the hand-trajectory tokens $\mathbf{E}_h$ are used as keys K and values V. This allows the model to enrich the video representation with motion patterns of the camera wearer's hands, such as reaching, grasping, moving, or releasing objects, which are often crucial for localizing manipulation-related events.

2.3 Query-Conditional Fusion

The Query-Conditional Fusion module combines the outputs of the two bimodal encoders, $\mathbf{E}_{vt}$ and $\mathbf{E}_{vh}$, into a single stream of video-level token embeddings $\mathbf{E}_o \in \mathbb{R}^{D \times T_v}$. Instead of fusing both branches uniformly, the module uses a query-conditional weighting mechanism that assigns a scalar weight to each branch according to the content of the query. The motivation is to exploit hand-trajectory information primarily when it is relevant, such as for queries involving hand–object interactions, while reducing its influence when it is uninformative or potentially misleading.

Formally, the query embedding is first aggregated via mean pooling:

$$\bar{\mathbf{e}}_q = \text{mean-pool}(\mathbf{E_t}) \in \mathbb{R}^D. \tag{1}$$

Then, the resulting vector is projected through a two-layer MLP with a GELU non-linearity to produce K=2 weighting logits:

$$\boldsymbol{\ell} = \mathbf{W}_2\, \text{GELU}(\mathbf{W}_1\, \bar{\mathbf{e}}_q) \in \mathbb{R}^K. \tag{2}$$

Next, a softmax operation converts the logits into a probability distribution over the two branches:

$$w_k = \frac{\exp(\ell_k)}{\sum_{j=1}^{K} \exp(\ell_j)}. \tag{3}$$

Finally, the fused representation $\mathbf{E_o}$ is the weighted sum of both branch outputs:

$$\mathbf{E_o} = w_1\, \mathbf{E_{vh}} + w_2\, \mathbf{E_{vt}}. \tag{4}$$

2.4 Temporal Segment Prediction

This module uses a multi-scale temporal feature pyramid [5] to localize answer spans of different durations. Starting from the fused video sequence of tokens $\mathbf{E}_o$, a stack of temporal self-attention Transformer blocks progressively aggregates contextual information and downsamples the temporal resolution, producing a set of feature levels with different temporal granularities. High-resolution levels preserve fine temporal details that are useful for short answer spans, whereas lower-resolution levels capture longer-range context and are better suited for extended events.

For each level of the temporal pyramid, the prediction head evaluates candidate temporal locations and estimates their corresponding answer boundaries. Concretely, two lightweight 1D convolutional heads are applied: a classification head predicts the probability that each temporal location belongs to a query-relevant foreground segment, while a regression head predicts the distances from that location to the start and end boundaries of the answer span to further refine the final temporal boundaries. The final prediction is obtained by selecting the highest-scoring temporal candidates and converting their regressed boundary offsets into the temporal interval $[t_s, t_e]$.

2.5 Training

The proposed model is trained with the grounding loss $\mathcal{L}_{\text{grd}}$. In particular, the temporal prediction module produces, for each level of the multi-scale temporal pyramid, foreground/background confidence scores and boundary offsets for candidate temporal locations. The grounding loss therefore combines a binary classification loss, which supervises whether each candidate belongs to the answer segment, and an IoU-based regression loss, which supervises the predicted start and end boundaries:

$$\mathcal{L}_{\text{grd}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{reg}}^{\text{IoU}}. \tag{5}$$

The following training strategy is adopted [5]: AdamW with a base learning rate $\mathrm{lr}_{\mathrm{base}} = 5 \times 10^{-5}$, cosine learning-rate decay, two warmup epochs, ten training epochs, and a batch size of 2. The newly introduced modules, namely the bimodal encoders and the Query-Conditional Fusion module, are trained with a learning-rate multiplier of 2. The pretrained unimodal encoders remain frozen, except for the trajectory encoder, which is trained jointly with the rest of the proposed modules.

3 Experiments

3.1 Dataset

We evaluate our method on the dataset Ego4D NLQ v2 [4], which contains 13,781 training and 4,536 validation query–clip pairs. Models are trained on the training split and evaluated on the validation split using the standard Rm@IoU=n metric, which measures the percentage of queries for which at least one of the top-m predicted moments achieves an IoU greater than or equal to n with the ground truth. Notice that the evaluation split is strictly used as a test set and therefore it has not been used for the model training. We further group queries into four template-derived categories: Hand–Object Interaction (HOI, N=1,929 in validation), Object Location (ObjLoc, N=1,661), People (N=228), and Quantity/State (N=718). Figure 2 summarizes the query distribution in Ego4D NLQ v2 across categories and splits. The dataset shows a clear imbalance, with Hand–Object Interaction and Object Location being the most frequent query types.

3.2 Ablation Study

We compare three model variants to isolate the contribution of each introduced component:

- **Baseline**: GroundNLQ [5] without the trajectory modality. The model relies only on the visual–text representation, which is directly passed to the prediction head.
- **Uniform fusion**: We extend the baseline by enabling the trajectory encoder and introducing the two bimodal encoders, but combine their outputs with equal weights, independently of the query. This variant removes query-dependent weighting and serves as a query-agnostic fusion baseline, allowing us to isolate the contribution of the query-conditional weighting mechanism.
- **Query-Conditional Fusion** (**Ours**): Full model of Sec. 2, where the two bimodal encoders are combined through a query-conditional weighting mechanism.

The results are shown in Table 1. Query-Conditional Fusion achieves the best overall performance. Comparing Uniform fusion with the baseline reveals that hand trajectory is an asymmetrically useful modality: it improves HOI queries

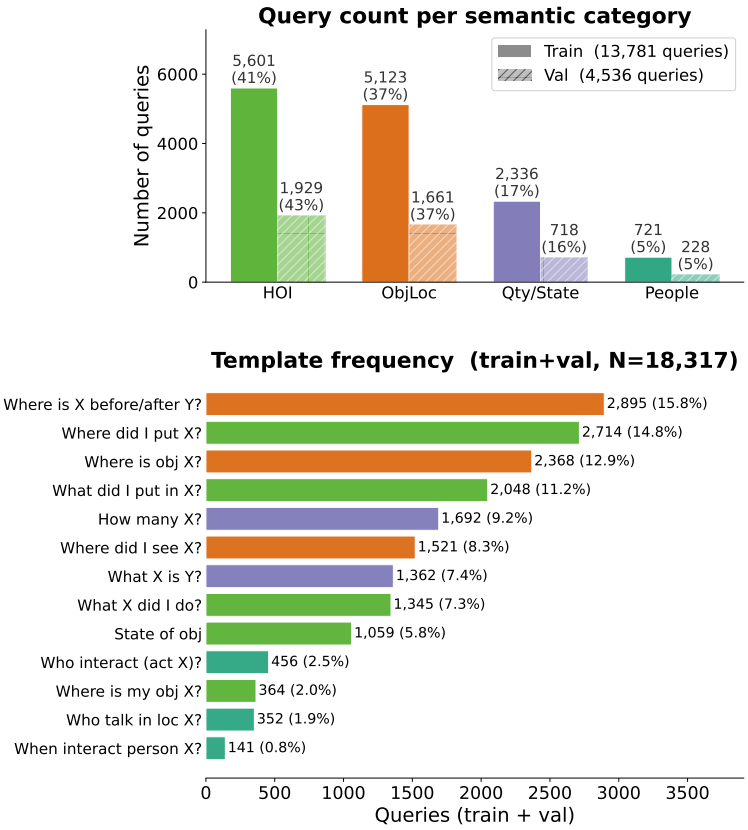

Fig. 2. Query distribution on Ego4D NLQ v2. **Top**: per-category counts for the training and validation splits, where Hand–Object Interaction and Object Location dominate and People queries remain rare. **Bottom**: per-template query frequency (train+val), color-coded by category. Templates are grouped into four semantic categories with consistent train/val distributions.

by +2.55 but introduces noise on queries where kinematics are irrelevant (performance degradation for People of −1.32 and for Object Location of −2.53). Query-Conditional Fusion mitigates this asymmetry, recovering the loss on People (+2.19 over Uniform fusion, +0.87 over the baseline) and largely closing the gap on Object Location (+1.38 over Uniform fusion), while preserving the HOI gains. This indicates that the weighting mechanism learns to modulate the hand-motion branch when motion cues are not discriminative for the query.

Overall, this leads to improvements or matched performance across all categories, with gains of +0.57 over Uniform fusion and +1.34 over the baseline. Fig. 3 illustrates this behavior on representative clips. For each category, we show three frames around the predicted moment (before, peak, after), overlaid with the detected hand skeleton and its trajectory. Note that hand skeletons are not

Table 1. Per-category R1@0.3 on Ego4D NLQ v2.

Category	N	Baseline	Uniform fusion	Query-Conditional Fusion
Hand–Object Interaction	1929	28.99	31.54	31.65
Object Location	1661	22.34	19.81	21.19
People	228	26.32	25.00	27.19
Quantity / State	718	24.93	29.25	29.25
Overall	4536	25.77	26.54	**27.11**

detected in every frame, since hands may be occluded or out of the field of view. The qualitative results highlight the same pattern observed quantitatively. In HOI and Quantity/State queries, the trajectory is strongly concentrated around the interaction moment, providing a reliable temporal cue. In contrast, for Object Location and People queries, hand kinematics are sparse or unrelated to the queried entity, which explains why Query-Conditional Fusion is necessary to prevent the model from being misled by irrelevant motion cues.

4 Conclusion

We introduced hand trajectory as an explicit auxiliary modality for egocentric Natural Language Query grounding and proposed a Query-Conditional Fusion mechanism to exploit it selectively. The proposed approach encodes hand skeleton trajectories as video-aligned kinematic features and combines them with visual and textual representations through a query-dependent weighting module. This allows the model to increase the influence of hand-motion cues for manipulation-related queries while reducing their contribution when they are uninformative or potentially misleading.

Experiments on Ego4D NLQ v2 show that hand trajectory is an asymmetrically useful modality: it improves Hand–Object Interaction queries when incorporated into the model, but can harm categories such as People and Object Location if fused uniformly. Query-Conditional Fusion addresses this issue by adapting the contribution of each branch to the query content, achieving the best overall performance and improving or matching over uniform fusion across all query categories. The qualitative results further confirm that hand trajectories provide reliable temporal cues for interaction-centric queries, while query-adaptive fusion prevents irrelevant motion from degrading localization in other cases.

Overall, our results suggest that egocentric NLQ grounding benefits not only from adding new auxiliary modalities, but also from deciding when each modality should be trusted. This opens future directions toward richer query-aware fusion strategies that incorporate hand–object relations, object dynamics, and other egocentric cues in a more adaptive manner.

References

1. Feng, Y., Zhang, H., Liu, M., Guan, W., Nie, L.: Object-shot enhanced grounding network for egocentric video. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24190–24200 (2025)
2. Feng, Y., Zhang, H., Xie, Y., Li, Z., Liu, M., Nie, L.: Objectnlq@ ego4d episodic memory challenge 2024. arXiv preprint arXiv:2406.15778 (2024)
3. Furnari, A., Farinella, G.M.: What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In: Proceedings of the IEEE/CVF International conference on computer vision. pp. 6252–6261 (2019)
4. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18995–19012 (2022)
5. Hou, Z., Ji, L., Gao, D., Zhong, W., Yan, K., Li, C., Chan, W.K., Ngo, C.W., Duan, N., Shou, M.Z.: Groundnlq@ ego4d natural language queries challenge 2023. arXiv preprint arXiv:2306.15255 (2023)
6. Lin, K.Q., Wang, J., Soldan, M., Wray, M., Yan, R., Xu, E.Z., Gao, D., Tu, R.C., Zhao, W., Kong, W., et al.: Egocentric video-language pretraining. Advances in Neural Information Processing Systems **35**, 7575–7586 (2022)
7. Lin, K.Q., Zhang, P., Chen, J., Pramanick, S., Gao, D., Wang, A.J., Yan, R., Shou, M.Z.: Univtg: Towards unified video-language temporal grounding. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2794–2804 (2023)
8. Lin, W.C., Lien, C.M., Lo, C., Yeh, C.H.: Gazenlq @ ego4d natural language queries challenge 2025 (2025)
9. Liu, Y., He, J., Li, W., Kim, J., Wei, D., Pfister, H., Chen, C.W.: R 2-tuning: Efficient image-to-video transfer learning for video temporal grounding. In: European conference on computer vision. pp. 421–438. Springer (2024)
10. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
11. Shan, D., Geng, J., Shu, M., Fouhey, D.F.: Understanding human hands in contact at internet scale. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9869–9878 (2020)
12. Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., et al.: Internvideo: General video foundation models via generative and discriminative learning. arXiv preprint arXiv:2212.03191 (2022)
13. Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C.L., Grundmann, M.: Mediapipe hands: On-device real-time hand tracking. arXiv preprint arXiv:2006.10214 (2020)
14. Zhong, E., del Blanco, C.R., Berjón, D., Jaureguizar, F., García, N.: Real-time monocular skeleton-based hand gesture recognition using 3d-jointsformer. Sensors **23**(16) (2023)

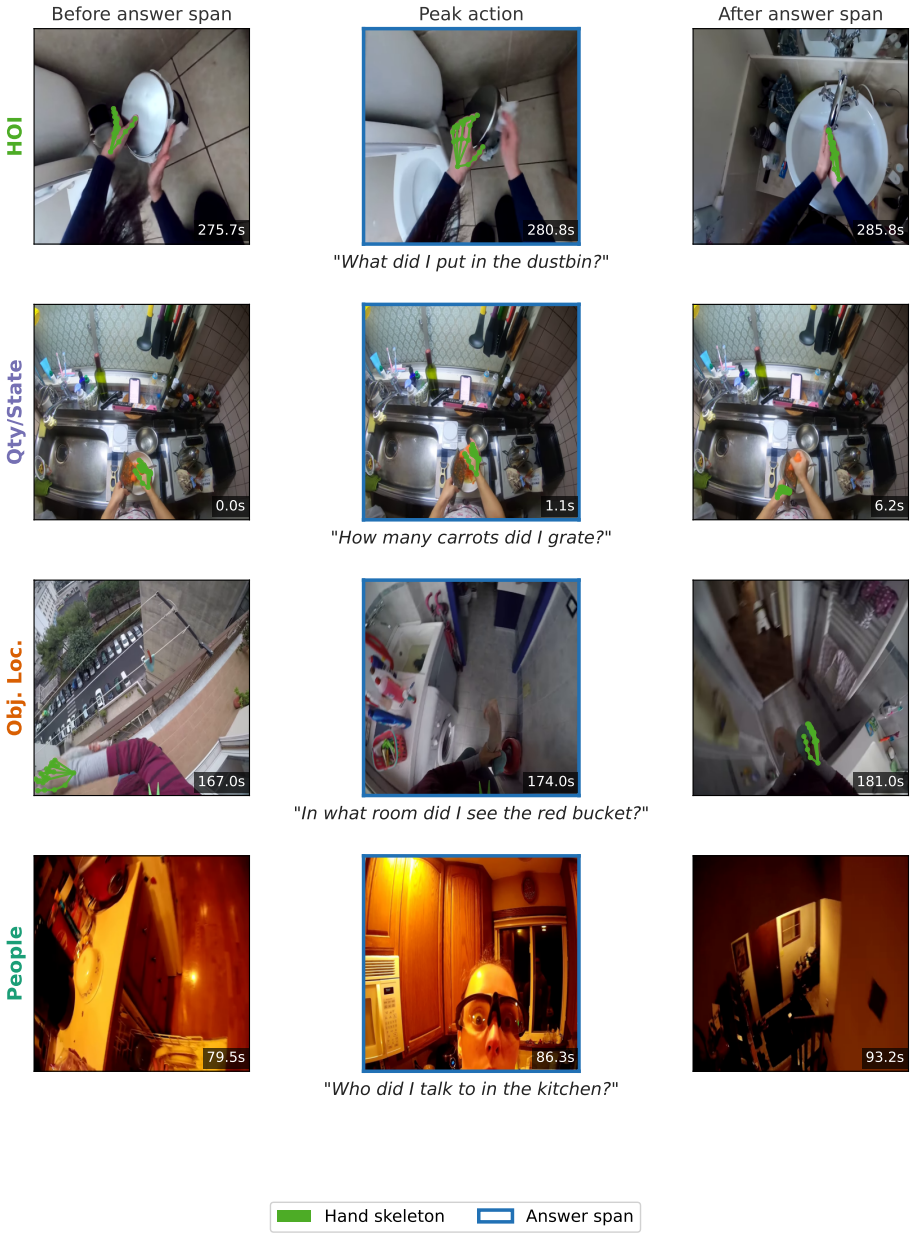

Fig. 3. Qualitative examples on Ego4D NLQ v2 validation. One row per query category (top to bottom: HOI, Quantity/State, Object Location, People), with three frames sampled around the ground-truth span (before, peak, after).