# HERBench: A Benchmark for Multi-Evidence Integration in Video Question Answering

## Supplementary Material

## 1. Implementation Details

This section provides comprehensive implementation details for the HERBench construction pipeline, which employs a tripartite structure to enforce high evidential requirements (ER). We detail the algorithms, mathematical formulations, thresholds, and quality control procedures used to transform raw videos into the final dataset.

### 1.1. Track Ranking and Selection

**Tracking and Trajectory Refinement.** To ensure robust performance in our benchmark's highly dynamic environments—characterized by dense crowds and significant depth variations—we conducted a rigorous empirical evaluation of state-of-the-art detection and tracking pipelines. Specifically, we benchmarked three leading detectors (*RF-DETR-L* [19], *YOLO-v11-x* [7], and *Grounding-DINO* [11]) in combination with two widely adopted multi-object trackers (*DeepSORT* [25] and *ByteTrack* [29]). This evaluation was performed on a curated subset of the DanceTrack [22] benchmark, explicitly selected to mirror the crowd density, complex motion dynamics, and high inter-subject appearance diversity typical of our video corpus.

As detailed in Table 1, the combination of *RF-DETR-L* [19] and *DeepSORT* [25] yielded the highest tracking fidelity, achieving a HOTA of 47.1% and an IDF1 of 53.2%. Consequently, we adopted this configuration as the foundation of our pipeline. We utilize this detection-tracking stack by applying a high-recall *RF-DETR* [19] detector with a confidence threshold of 0.3 and a per-frame cap of 300 detections. Association within *DeepSORT* [25] uses a two-stage IoU matching: high-confidence detections (score $> 0.5$) are matched with an IoU threshold of 0.7, followed by lower-confidence detections with a relaxed IoU threshold of 0.35.

| HOTA / IDF1 (%) | RF-DETR-L | YOLO-v11-x | Grounding-DINO |
|---|---|---|---|
| **DeepSORT** | **47.1 / 53.2** | 43.3 / 51.1 | 42.9 / 50.6 |
| **ByteTracker** | 46.9 / 52.7 | 43.3 / 51.0 | 42.6 / 50.5 |

Table 1. Tracking pipeline benchmarking on a curated DanceTrack [22] subset mirroring HERBench's dynamic characteristics. The combination of RF-DETR-L and DeepSORT achieved the highest scores and was selected for our data generation pipeline.

To enforce physical plausibility, we apply an *outlier removal* step that explicitly discards per-frame boxes implying implausible motion (velocity $> 50$ pixels/frame) to eliminate spurious detections. To ensure continuity, we apply gap interpolation for missing detections up to 30 frames (1s at 30 fps) and trajectory smoothing via Gaussian filtering (window size 5). We specifically address track fragmentation by detecting merge candidates $(T_i, T_j)$ that are temporally ordered with a gap $\leq 30$ frames and spatially compatible. We minimize the following merge cost:

$$C_{merge} = \Delta t_{gap} + \frac{\|c^i_{last} - c^j_{first}\|_2}{\text{IoU}(box^i_{last}, box^j_{first})} \quad (1)$$

where $c$ denotes the bounding box centroid. The overall tracking, post-processing, and ranking pipeline is visualized in Figure 1.

**TrackRank scoring function.** To select the top $m \in [6, 10]$ salient entities per video, we compute a composite *TrackRank* score $S_i$ that aggregates metrics for each track $i$ (all computed per video and normalized by the maximum over tracks). Unlike simple duration-based ranking, we use the following weighted formulation:

$$S_i = \frac{\sum_k w_k \cdot M_{i,k}}{\sum_k w_k} \quad (2)$$

The specific components and their empirically tuned weights are:

- **Duration** ($w = 2.0$) **& Size** ($w = 1.0$): Favors tracks with sustained presence and higher average bounding box area.
- **Associated Objects** ($w = 2.0$): Normalized count of distinct non-person object classes overlapping the person's box (IoU $> 0.2$).
- **Center Distance** ($w = 2.4$) **& Motion** ($w = 1.0$): Euclidean distance between first and last centroids, favoring traversals over stationary behavior.
- **Appearance Exceptionality** ($w = 2.2$): We quantify rarity as the normalized L1 distance from the dataset's average appearance in feature space (HSV and LBP histograms).
- **Scene Coverage** ($w = 1.5$): Area of the Convex Hull enclosing the track's boxes.
- **Quality Metrics**: Aggregates *Average Confidence* ($w = 0.8$, mean detection score), *Smoothness* ($w = 0.7$, computed as 1 minus normalized acceleration magnitude to penalize jitter), and *Aspect-Ratio Stability* ($w = 0.5$, defined as 1 minus the standard deviation of width/height ratios to penalize shape fluctuations).

**Hard Filter Cascade.** Prior to ranking, we enforce a hard filter: we keep only the COCO [10] "person" class, require length $\geq 20$ frames, average area $\geq 5,500$ pixels, and require the track center to fall within the central safe region (frame cropped by 10% margins) in at least 5 frames.

**Diversity Sampling Strategy.** To ensure diversity among the selected tracks, we employ a round-robin selection across rankings generated from multiple perturbed weight configurations ($\gamma \sim U(0.5, 1.5)$). This prevents redundancy (e.g., selecting visually identical pedestrians) and ensures a broad coverage of high-quality entities, which are subsequently manually validated to exclude phantom detections or identity switches.

**Track Selection as Noise Control.** Per video, we select the top 6–10 tracks according to the TrackRank composite score (Eq. 2). Post-hoc analysis of the selected tracks' average detection confidence shows that they consistently fall within the top $\sim$20% most confident tracks per video, confirming that the ranking procedure implicitly filters out low-confidence, noise-prone trajectories before they can propagate errors into the question generation pipeline. The TrackRank selection process thus doubles as a noise control mechanism: by funneling only high-confidence, well-resolved trajectories into downstream task programming, it substantially mitigates the risk of identity switches, fragmented detections, or phantom tracks propagating into the ground-truth labels.

## 1.2. Decoupled Descriptor Generation

**A-card and B-card generation.** For each selected track, we generate disentangled descriptions using GPT-4o [15]. We sample 10-11 crops, reserving the first and last 20% of the trajectory for Appearance (A-cards) and the middle 60% for Behavior (B-cards). This ensures a temporal gap of at least 30 frames between appearance and behavior cues. An example of the resulting disentangled A- and B-cards for a single track is shown in Figure 2. We use the following prompt structure:

**System prompt.** For the following tasks, use only your vision capabilities. When referring to directions, use the camera's point of view.
**1. Person Description.** All images depict the same individual. In 2–4 sentences, describe their appearance in detail: clothing types and colors, accessories, hair, body build, and any distinctive features that make them easy to pick out. *Do not mention position in the frame or any actions.*
**2. Path Description.** In 3–7 sentences, describe the person's path and behavior over time. Mention the overall path shape, entry and exit edges, stops, and interactions. *Do not repeat any appearance details from the first description.*

To visualize the output of this pipeline, Figure 2 presents qualitative examples of the generated Appearance (A) and Behavior (B) cards alongside their corresponding tracked image crops. These examples highlights the effectiveness of the temporal split: the tracked visual crops from the start and end of the trajectory inform the static attribute descriptions in the A-card, while the central frames drive the dynamic action summaries in the B-card. This separation ensures that the descriptors remain disentangled.

**Leakage prevention.** To strictly enforce the "Look & Separate" principle, we calculate the token-level Jaccard similarity between the generated A-card and B-card. We set the Jaccard threshold to 0.15 based on manual inspection: above this, descriptors often share explicit appearance/behavior leakage.

## 1.3. Spatial Operations and Region Definitions

**Entry/exit edge labeling.** For tasks like *Region-Localized People Counting (RLPC)*, we define entry and exit edges based on the position of a track's centroid in its first and last frames. Let $c_t = (x_t, y_t)$ be the centroid at frame $t$ of a track with start frame $t_{\text{start}}$ and end frame $t_{\text{end}}$, and let $W, H$ denote the frame width and height. We say that a track enters through edge $e$ if $c_{t_{\text{start}}}$ lies in the corresponding edge band, and exits through edge $e'$ if $c_{t_{\text{end}}}$ lies in the band of $e'$. The top edge band is defined as $y < 0.3H$, the bottom as $y > 0.85H$, and the left/right edges as the outer 15% of the width ($x < 0.15W$ and $x > 0.85W$, respectively).

**Region-of-interest (ROI) membership.** For *[RLPC]*, we also define rectangular ROIs (e.g., frame halves or specific zones). A track is counted as visiting an ROI if, at any frame, at least 50% of its bounding box area lies within the region (Intersection-Over-Box $\geq 0.5$). We count the unique track IDs that satisfy this predicate to derive people counts under spatial constraints. To absorb residual tracking noise (missed detections, fragmented tracks), multiple-choice options are reported as binned count ranges rather than exact integers. The bins are constructed so that the correct range spans approximately ±40% around the true count on average, ensuring that minor tracking errors do not invalidate the ground truth while still requiring models to perform meaningful spatial counting, see Figure 16 for example.

**Duration computation (*[MPDR]*).** We compute visible-time intervals $(t_{start}, t_{end})$ for every track. Using interval algebra, we determine ground truth for questions such as "Who stayed longest?" or "Who entered first?" by comparing duration scalars $(t_{end} - t_{start})$ and timestamps.

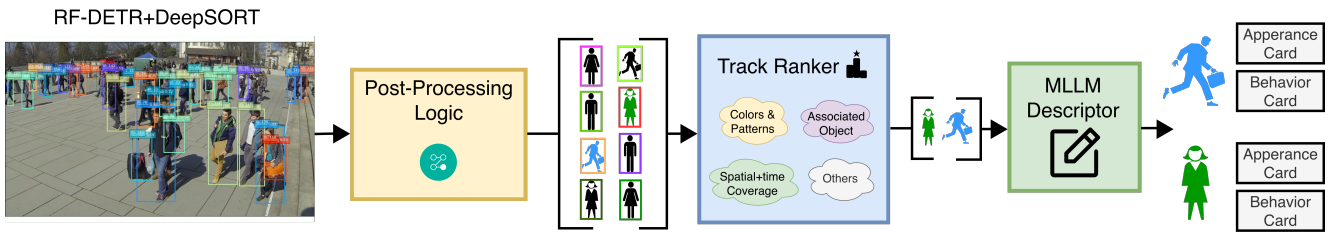



Figure 1. **Tracking, post-processing, and ranking pipeline.** RF-DETR [19] detections are linked with DeepSORT [25] into raw person tracks, followed by outlier removal, gap interpolation, and Gaussian smoothing. A TrackRanker then scores and selects salient trajectories, which are passed to an MLLM descriptor module to generate temporally decoupled appearance (A) and behavior (B) cards that serve as the scaffold for downstream HERBench tasks.

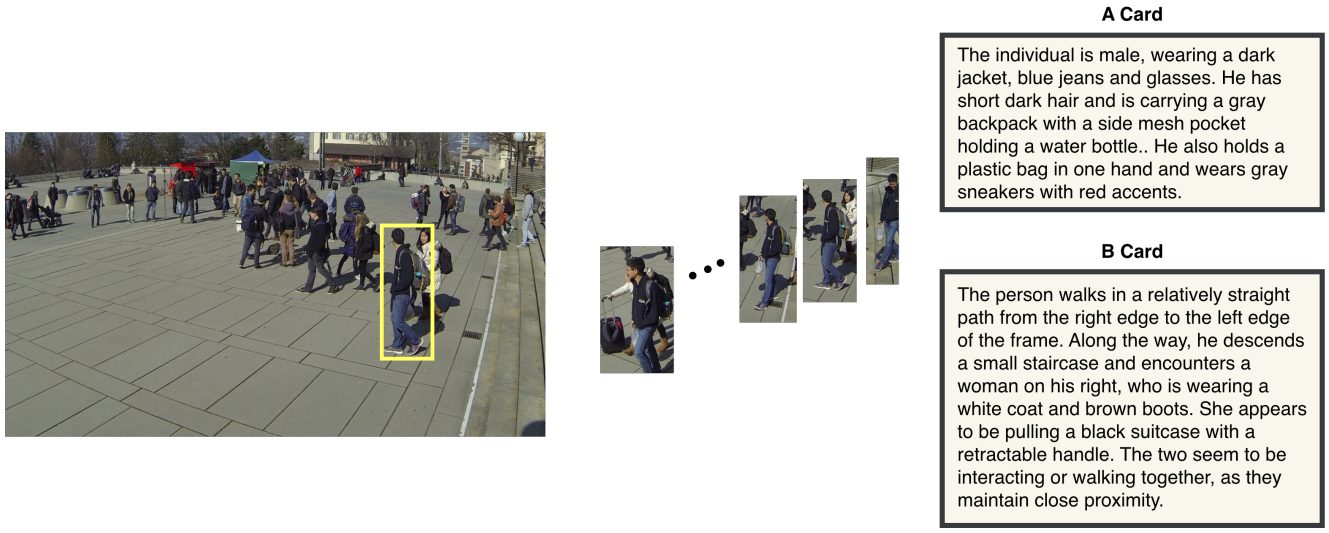



Figure 2. **Example of disentangled A- and B-cards.** For a single tracked individual (highlighted trajectory in the top-left strip), we show the sampled frames and the corresponding appearance (A-card) and behavior (B-card) descriptions. The A-card captures only static visual attributes (clothing, colors, accessories, physique), while the B-card describes the person's path, timing, and interactions over time without repeating appearance cues, enforcing the "Look & Separate" principle.

### 1.4. Scene Card Perturbations

**Shot Segmentation and Description.** We use *TransNetV2* [21] for shot boundary detection. To calibrate its reliability on our video corpus, we manually reviewed shot boundaries on 30 videos used for *[TSO]* and *[SVA]* tasks (34% of videos contributing to these tasks). Comparing TransNetV2 [21] predictions against manual annotations yields F1 = 0.97, confirming that shot segmentation noise is negligible. For the *[SVA]* task, faithful scene cards are generated via an MLLM using the following prompt:

> *"Describe concisely the scene in one sentence without reference to the 'scene', refer (if relevant) to the entities, genders and appearance (type and colors of hair/clothing/accessories) of each entity, occurrence, actions, background, and location."*

**Perturbation Engine.** To generate negative samples for *[SVA]*, we prompt the model to modify faithful descriptions by altering 2-5 atomic details. The prompt constraints ensure:

- **Modifications**: Change existing details (color, count, attributes).
- **Additions**: Insert plausible but absent elements (extra objects, background items).
- **Plausibility**: Changes must be false but highly plausible within the context of the video.

An example of a faithful scene card and its perturbed counterpart used for the *[SVA]* task is shown in Figure 3.

### 1.5. Corpus-Plausible Foil Generation

**Ground Truth Integration.** For tasks requiring verification of absence, we leverage human-verified event logs.

- **False Action Memory (*[FAM]*)**: We sample a "false" action by pairing an object present in the video with an ac-

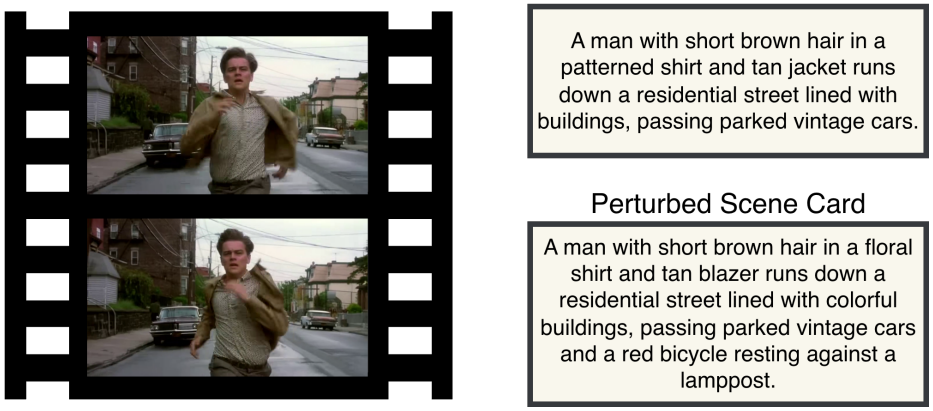



Figure 3. **Faithful and perturbed scene cards for SVA.** The top card provides a faithful one-sentence description of a shot, mentioning the main actor, appearance, background, and motion. The bottom card is a perturbed variant where 2-5 atomic details (e.g., clothing pattern, background appearance, additional objects) are modified or added while remaining globally plausible. These pairs form positive and negative options in the Scene Verification & Arrangement task, probing fine-grained scene-level sensitivity to small but visually significant details.

tion from the corpus that does *not* occur in the current video.

- **False Object Memory (*[FOM]*)**: We select an absent object from the corpus-wide index that is compatible with actions present in the video (e.g., if "cutting" occurs, "carrot" is a valid distractor if absent).
- **Action Counting (*[AC]*)**: Distractor counts are generated such that the correct count's rank varies uniformly across options.
- **Action Sequence Integrity (*[ASII]*)**: We sample a 5-event ground-truth timeline. Distractors are generated using two perturbation functions: `swap_mid` (swapping two non-adjacent events) and `rotate` (shifting the sequence). Crucially, we verify against the event log that the perturbed timeline does not accidentally exist in the video.

## 1.6. Text-Only Bias Filtering Details

**Filtering procedure.** To suppress language priors, we apply a rigorous Text-Only Filtering stage. We discard any question correctly answered by $\geq$ 3 of 4 blind LLMs (Qwen2-7B [17], Qwen2.5-7B [18], Llama-3-8B [14], and Vicuna-7B v1.5 [5]). This step rejects approximately 10% of candidates (e.g., questions answerable via object-color co-occurrence priors).

## 1.7. Human Verification Protocol

Experts conduct verification on a stratified 15% sample of instantiated questions to audit whether the construction pipeline preserves the intended evidential constraints. The audit focuses on three properties: (i) **minimum frame-set compliance**, i.e. confirming that the item requires at least three distinct frames; (ii) **uniqueness of the answer**, i.e. confirming the existence of a single objective ground-truth answer; and (iii) **descriptor disentanglement**, i.e. verifying that A-cards and B-cards do not leak information from one another. Items that violate any of these conditions are rejected. This process resulted in a 17.8% rejection rate.

| Task | Full video (%) | Oracle frames (%) |
|---|---|---|
| *[TSO]* | 91.7 | 93.8 |
| *[MPDR]* | 88.3 | 97.0 |
| *[ASII]* | 86.7 | 95.8 |
| *[AGBI]* | 95.8 | 98.1 |
| *[AGAR]* | 95.0 | 98.3 |
| *[AGLT]* | 92.5 | 97.4 |
| *[FAM]* | 84.2 | 92.8 |
| *[SVA]* | 84.2 | 94.8 |
| *[FOM]* | 87.5 | 96.3 |
| *[MEGL]* | 85.8 | 95.4 |
| *[AC]* | 84.2 | 97.7 |
| *[RLPC]* | 90.0 | 90.9 |
| **Overall** | **88.8** | **95.7** |

Table 2. **Per-task human accuracy.** Accuracy of human annotators in the full-video and oracle-frame settings across all HERBench tasks. In the full-video setting, annotators answer with unrestricted video access and free scrubbing. In the oracle-frame setting, annotators answer using only the curated oracle frame-set, without access to the source video.

## 1.8. Human Validation and Oracle-Frame Study

To complement the construction-time audit above, we conducted two human studies that assess HERBench from two complementary perspectives: overall question answerabil-

ity and oracle-evidence sufficiency. Each study used a separate group of 6 annotators.

**Study design.** In the **full-video** setting, annotators answered a shared set of 240 questions spanning all 12 tasks (20 questions per task) with unrestricted video access and free scrubbing. In the **oracle-frame** setting, annotators answered 2,160 questions using only the curated oracle frame-set provided by the benchmark construction pipeline, without access to the source video. Each oracle item was presented in the same format used by the oracle-based analysis in the main paper, namely the curated evidence frames together with distractor frames.

**Results.** Table 2 reports the per-task accuracies. In the full-video setting, annotators achieved 88.8% accuracy overall, with substantial inter-annotator agreement (Fleiss' $\kappa = 0.74$), indicating that the benchmark remains highly answerable for humans despite its high evidential demand. In the oracle-frame setting, annotators achieved 95.7% accuracy overall, showing that the curated oracle frame-set is generally sufficient to resolve the question without access to the full temporal context. The largest improvements appear in tasks such as *[AC]*, *[SVA]*, and *[MEGL]*, where the answer depends on sparse or temporally localized evidence.

**Benchmark cleanup.** After each oracle-frame response, annotators were shown the ground-truth answer and invited to flag problematic items. This process surfaced three types of issues: mis-indexed evidence frames, incorrect ground-truth labels, and genuinely ambiguous items. Specifically, annotators flagged 42/2160 items (1.9%) for evidence mis-indexing, 18/2160 (0.8%) for incorrect ground truth, and 58/2160 (2.7%) as ambiguous. All flagged items were subsequently corrected or removed from the final release.

### 1.9. Dataset Statistics

**Scale and Video Characteristics.** HERBench comprises 26,806 questions derived from 336 unique videos. The videos feature substantial duration (avg. 395s, range 60-2100s) to ensure temporal dispersion of evidence. Sources include HD-EPIC [16], WildTrack [4], PersonPath22 [20], and movie trailers.

**Question Properties.** The average question length is 65.5 tokens with a vocabulary of ∼7.3k unique word types. Questions are strictly balanced across 5 multiple-choice options. The mean temporal span of evidence required per question is 101.1 seconds.

## 2. Extended Experimental Results & Analysis

We provide a deeper quantitative analysis of the challenges posed by HERBench, expanding on the MRFS metrics and frame selection ablation.

### 2.1. Extended MRFS Analysis

**Per-Task MRFS.** Table 3 details the Minimum Required Frame-Set statistics. We observe a distinct correlation between the reasoning scope of a task and its evidential requirement. Tasks requiring global chronology and the integration of multiple semantic units, specifically *[TSO]* (Temporal Shot Ordering, MRFS 9.05), *[FAM]* (False Action Memory, MRFS 6.77), and *[SVA]* (Scene Verification, MRFS 6.74), naturally exhibit the highest MRFS. To answer these questions correctly, a model must aggregate evidence from widely dispersed video segments or perform an exhaustive search to verify absence, effectively precluding single-frame shortcuts.

In contrast, tasks focused on local attributes or spatially constrained counting, such as *[RLPC]* (Region-Localized People Counting, MRFS 3.11) and *[AGAR]* (Attribute Recognition, MRFS 3.85), require fewer distinct frames. However, even these "lower" MRFS values demonstrate that reliance on a single frame is insufficient, confirming that HERBench successfully enforces multi-evidence integration even for localized tasks. The overall weighted mean MRFS of 5.49 validates the benchmark's design goal: forcing models to look at multiple snapshots to derive correct answers.

| Task | Total | Mean MRFS |
|---|---|---|
| *[TSO]* | 2123 | **9.05** |
| *[MPDR]* | 2717 | 4.30 |
| *[ASII]* | 2127 | 6.00 |
| *[AGBI]* | 1226 | 3.81 |
| *[AGAR]* | 876 | 3.85 |
| *[AGLT]* | 2362 | 4.45 |
| *[FAM]* | 1962 | 6.77 |
| *[SVA]* | 4569 | 6.74 |
| *[FOM]* | 2022 | 5.14 |
| *[MEGL]* | 3061 | 6.33 |
| *[AC]* | 1623 | 5.26 |
| *[RLPC]* | 2138 | 3.11 |
| **Total / Weighted Mean** | **26,806** | **5.49** |

Table 3. **Per-task MRFS statistics** Computed with $x = 16$ using Qwen2.5-VL [3] and AKS [23].

**MRFS vs Accuracy** As illustrated in Figure 4, there is a pronounced inverse relationship between the evidential demand of a benchmark—quantified by the Mean MRFS—and the performance of state-of-the-art Video-LLMs. Benchmarks with low evidential requirements,

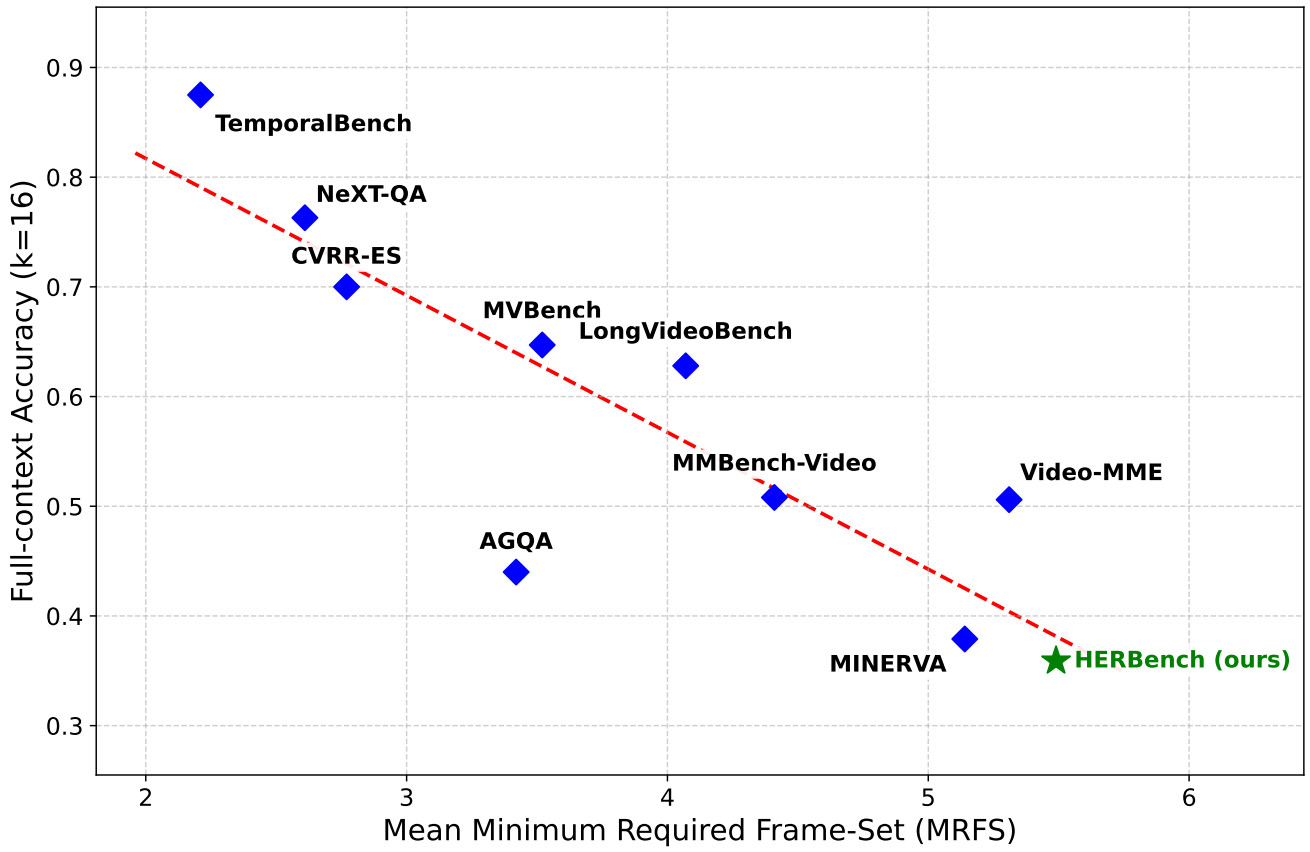



Figure 4. **Impact of Evidential Requirement on Model Accuracy.** We plot the Mean Minimum Required Frame-Set (MRFS) against Full-context Accuracy ($k = 16$), measured using Qwen 2.5 VL 7B [3], across ten video QA benchmarks. The dashed line indicates the fitted trend. A clear inverse correlation emerges: as the evidential burden increases (higher MRFS), model performance tends to decrease ($R^2 = 0.82$). HERBench (green star) occupies the high-demand end of the spectrum (MRFS 5.49), highlighting the challenges current Video-LLMs face in multi-evidence integration relative to lower-requirement benchmarks such as TemporalBench and NeXT-QA.

such as TemporalBench (2.21 MRFS, 87.5%) and NeXT-QA [27] (2.61 MRFS, 76.3%), allow Qwen 2.5 VL 7B [3] to achieve relatively high accuracy, possibly due to the feasibility of single-frame shortcuts or language priors. Mid-range benchmarks—CVRR-ES, MVBench, LongVideoBench, and MMBench-Video—cluster between 2.8–4.4 MRFS with accuracies that progressively decline from 70.0% to 50.8%, tracing the fitted trend closely. At the high-demand end, MINERVA (5.14 MRFS, 37.9%), Video-MME (5.31 MRFS, 50.6%), and HERBench (5.49 MRFS, 35.9%) impose substantially greater evidential burdens, coinciding with markedly lower accuracies. Notably, AGQA (3.42 MRFS, 44.0%) falls below the trend line, suggesting that factors beyond evidential density—such as compositional question complexity—can independently depress performance. The broadened comparison across ten benchmarks strengthens the evidence for a *fusion deficit* in current architectures: while models may be effective at retrieving isolated frames, their capacity for compositional reasoning degrades consistently as the number of required evidence pieces grows. HERBench, positioned at the extreme of this spectrum, is specifically designed to stress-test this bottleneck by requiring the integration of non-redundant, temporally dispersed cues.

## 2.2. Full Frame-Selection Ablation

To more precisely disentangle the role of evidence retrieval from that of multi-evidence fusion, we perform an extensive ablation over five frame selection strategies—Uniform, Vanilla-BLIP [8], BOLT-ITS [12], AKS [23], and Oracle Frames (OF)—and evaluate their effect across all twelve HERBench tasks (Table 4). Overall, learned strategies such as BOLT-ITS and AKS provide moderate gains over Uniform sampling, reflecting their ability to prioritize query-relevant frames while maintaining broader temporal coverage. However, their improvements are uneven across tasks: both methods show the largest benefits in sparse-evidence settings such as *[TSO]* and *[FAM]*, where the critical ev-

Table 4. **Frame Selection Ablation.** Accuracy (%) on a random subsample of questions using InternVL3.5 [24], Qwen3-VL [2], and Ovis-2.5 [13] with Uniform, Vanilla-BLIP [8], BOLT-ITS [12], AKS [23], and GT Frames (OF) selectors. GT Frames represents the upper bound with manually curated evidence.

| Model | Selector | AC | AGAR | AGBI | AGLT | ASII | FAM | FOM | MEGL | MPDR | RLPC | SVA | TSO | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **InternVL3.5** | Uniform | 23.0 | 75.0 | 77.0 | 70.0 | 32.0 | 30.0 | 30.0 | 34.0 | 48.0 | 27.0 | 23.0 | 41.0 | 42.7 |
| | Vanilla-BLIP | 26.0 | 74.0 | 76.0 | 71.0 | 27.0 | 27.0 | 29.0 | 28.0 | 46.0 | 33.0 | 41.0 | 43.0 | 42.1 |
| | BOLT-ITS | 22.0 | 72.0 | 74.0 | 71.0 | 20.0 | 27.0 | 33.0 | 30.0 | 48.0 | 33.0 | 27.0 | 36.0 | 41.1 |
| | AKS | 27.0 | 66.0 | 77.0 | 74.0 | 36.0 | 29.0 | 30.0 | 35.0 | 54.0 | 33.0 | 33.0 | 17.0 | 42.7 |
| | **GT Frames** | 24.0 | 81.0 | 81.0 | 79.0 | 20.0 | 50.0 | 39.0 | 27.0 | 52.0 | 32.0 | 37.0 | 53.0 | **47.8** |
| **Qwen3-VL** | Uniform | 26.0 | 67.0 | 78.0 | 68.0 | 34.0 | 30.0 | 24.0 | 16.0 | 36.0 | 23.0 | 50.0 | 0.0 | 37.7 |
| | Vanilla-BLIP | 27.0 | 71.0 | 76.0 | 66.0 | 26.0 | 24.0 | 23.0 | 30.0 | 37.0 | 19.0 | 56.0 | 0.0 | 37.9 |
| | BOLT-ITS | 25.0 | 68.0 | 75.0 | 66.0 | 27.0 | 21.0 | 33.0 | 30.0 | 38.0 | 21.0 | 57.0 | 0.0 | 38.4 |
| | AKS | 24.0 | 65.0 | 73.0 | 69.0 | 29.0 | 22.0 | 27.0 | 20.0 | 35.0 | 22.0 | 49.0 | 0.0 | 36.2 |
| | **GT Frames** | 24.0 | 69.0 | 73.0 | 71.0 | 35.0 | 50.0 | 25.0 | 24.0 | 36.0 | 21.0 | 61.0 | 3.0 | **41.0** |
| **Ovis-2.5** | Uniform | 25.0 | 79.0 | 81.0 | 71.0 | 34.0 | 35.0 | 34.0 | 35.0 | 38.0 | 21.0 | 65.0 | 0.0 | 43.1 |
| | Vanilla-BLIP | 32.0 | 77.0 | 83.0 | 69.0 | 17.0 | 29.0 | 33.0 | 37.0 | 44.0 | 19.0 | 58.0 | 0.0 | 41.6 |
| | BOLT-ITS | 35.0 | 78.0 | 82.0 | 70.0 | 17.0 | 28.0 | 33.0 | 38.0 | 46.0 | 18.0 | 60.0 | 0.0 | 42.1 |
| | AKS | 25.0 | 76.0 | 80.0 | 74.0 | 31.0 | 39.0 | 39.0 | 30.0 | 49.0 | 17.0 | 51.0 | 0.0 | 42.6 |
| | **GT Frames** | 30.0 | 85.0 | 84.0 | 80.0 | 23.0 | 60.0 | 39.0 | 40.0 | 41.0 | 21.0 | 68.0 | 4.0 | **47.9** |

idence may appear only briefly within long videos. The oracle-based setting establishes an upper bound by supplying the manually curated evidence frames used during dataset construction. As shown in the rightmost column of Table 4, all three representative models experience nontrivial but still limited performance improvements in the OF regime (typically $+3$-6 absolute accuracy points relative to the best learned selector).

Importantly, the OF results highlight two key phenomena. First, even perfect access to the relevant frames does not resolve the majority of model failures: fusion-bound tasks such as *[AC]*, *[RLPC]*, and *[MEGL]* remain bottlenecks with accuracies barely above chance, indicating that retrieval is not the sole limiting factor. Second, improvements under OF are disproportionately large for temporally global tasks such as *[TSO]* and *[SVA]*, where correct reasoning requires coordinating multiple distant, non-overlapping visual clues. Here retrieval quality is a dominant factor, and learned selectors struggle to consistently surface all required frames. However, the inability of models to capitalize fully on oracle-quality evidence emphasizes that multi-frame integration itself remains a major unresolved challenge. Taken together, these results reinforce a two-stage deficit: (i) an *evidence retrieval bottleneck*, where existing selectors fail to reliably surface all critical cues, and (ii) a more fundamental *fusion bottleneck*, where models fail to combine available cues even when retrieval uncertainty is eliminated. HERBench's high evidential density and stringent cue separation make both deficits sharply visible, underscoring the need for future MLLMs to improve not only frame selection but also the downstream mechanisms for multi-cue aggregation.

### 2.3. MRFS robustness across backbones and frame selectors

To assess whether MRFS-based benchmark comparison is sensitive to the choice of backbone or selector, we evaluate MRFS under a range of configurations: four backbone models (Qwen2.5-VL [3], Gemini 2.5 Flash [6], LLaVA-OV-1.5 [1], and GPT-4o [15]) with AKS [23], and three selectors (AKS, BOLT [12], and $T^\star$ [28]) with Qwen2.5-VL (see Tab. 5). Here, $T^\star$ serves as a non-CLIP baseline. All results are computed on 50% stratified random samples from each benchmark. Across all tested settings, the benchmark ordering remains stable, with HERBench consistently yielding the highest MRFS, followed by LongVideoBench, MVBench, and NExT-QA. These results support MRFS as a robust measure of dataset-level evidential requirement.

## 3. Illustrative Examples for All Tasks

This section provides qualitative examples for all twelve HERBench tasks, each figure displays *one representative structured question* for the corresponding task. However, each task in HERBench contains *many distinct question structures and evidential templates*, and the examples below illustrate only a single instance of the broader variability present in the dataset.

**Temporal Reasoning & Chronology.** Figure 5 presents an example of the *Temporal Shot Ordering ([TSO])* task, which requires reconstructing the chronological order of

Table 5. MRFS robustness across models and keyframe selectors. Benchmarks: NExT-QA [27], MVBench [9], LongVideoBench [26], and HERBench.

| **Benchmark** | **Model on AKS (MRFS / Acc.)** | | | | **Selector on Qwen2.5-VL (MRFS / Acc.)** | | |
|---|---|---|---|---|---|---|---|
| | **Qwen2.5-VL** | **Gemini2.5F** | **LLaVA-OV-1.5** | **GPT-4o** | **AKS** | **BOLT** | $T^*$ |
| NExT-QA | 2.61 / 65.79 | 3.81 / 76.29 | 3.45 / 68.77 | 3.85 / 70.81 | 2.61 / 65.79 | 3.64 / 75.33 | 2.78 / 64.56 |
| MVBench | 3.52 / 56.71 | 3.92 / 57.41 | 3.69 / 54.24 | 4.09 / 54.42 | 3.52 / 56.71 | 3.73 / 56.28 | 3.53 / 55.53 |
| LongVideoBench | 4.07 / 41.38 | 4.57 / 64.59 | 4.50 / 61.49 | 4.92 / 59.01 | 4.07 / 41.38 | 4.89 / 49.14 | 4.11 / 48.59 |
| **HERBench** | **5.49 / 35.91** | **5.68 / 38.74** | **5.67 / 36.50** | **5.81 / 38.65** | **5.49 / 35.91** | **5.72 / 42.08** | **5.43 / 43.41** |

four non-overlapping shots. Figure 6 shows the *Multi-Person Duration Reasoning ([MPDR])* task, where models must compare visible-time intervals across multiple individuals. Figure 7 illustrates the *Action Sequence Integrity & Identification ([ASII])* task, requiring identification of the correct sequence among plausible permutations of narrated events.

**Referring & Tracking.** Figure 8 shows the *Appearance-Grounded Behavior Interactions ([AGBI])* task, where models must track a target described only by appearance and determine who interacts with them. Figure 9 provides an example of the *Appearance-Grounded Attribute Recognition ([AGAR])* task, requiring attribute extraction anchored to the tracked target. Figure 10 illustrates the *Appearance-Grounded Localization Trajectory ([AGLT])* task, where the model must infer how the target enters or exits the scene.

**Global Consistency & Verification.** Figure 11 presents the *False Action Memory ([FAM])* task, requiring verification of which plausible action did *not* occur in the video. Figure 12 shows the *Scene Verification & Arrangement ([SVA])* task, combining faithful and perturbed shot descriptions to assess fine-grained scene-level verification and ordering. Figure 13 depicts the *False Object Memory ([FOM])* task, requiring identification of a plausible but absent object interaction.

**Multi-Entity Aggregation & Numeracy.** Figure 14 provides an example of the *Multi-Entities Grounding & Localization ([MEGL])* task, where models must verify which appearance-described individuals actually appear in the video. Figure 15 illustrates the *Action Counting ([AC])* task, requiring enumeration of all instances of a specified action–object pair across the entire video. Finally, Figure 16 shows the *Region-Localized People Counting ([RLPC])* task, where the model must count unique individuals entering through specific spatial regions.

## [TSO] Temporal Shot Ordering

The following 4 shots (scenes) take place in the video:
1. A digital display shows the number "2" surrounded by illuminated control panels with labeled buttons, while a small hexagonal window reveals a reflection of a dark, futuristic suit amid glowing lights.
2. A view of Earth from space is overlaid with the text "Original Series August 12" shown against a backdrop of the planet's illuminated edge.
3. Two large, dark, organic-looking egg structures are positioned in a dimly lit room.
4. Reflections of two men, one holding a camera and the other in dark clothing, appear in a puddle on a street.

Select the option that correctly reflects the order in which these shots occur in the video:

A. 1->2->3->4

B. 4->1->3->2

C. 1->4->3->2

D. 1->4->2->3

E. At least two descriptions do not accurately reflect any shot from the video.

Figure 5

## [MPDR] Multi-Person Duration Reasoning

These people were in the video:
1. The individual is male, wearing a black puffer jacket and beige pants. He has a red backpack slung over one shoulder and dark shoes. His hair is short and he appears to be walking alongside another person.
2. The individual is male with short, dark hair. He is wearing a dark coat, a red scarf, and light gray pants. He accessorizes with glasses and carries a brown shoulder bag. His posture is slightly forward-leaning, and he appears to be focusing on a handheld device.
3. The individual appears to be male, wearing a dark navy jacket over polo shirt with white button and light blue jeans. He has short brown hair and is walking with a casual posture. He does not seem to carry any visible accessories.
Who stayed in the frame FOV for the longest time?

A. Person 3

B. Person 2

C. Person 1 and 3 both the same (less than 3 seconds difference)

D. Person 1

E. Person 1 and 2 both the same (less than 3 seconds difference)

Figure 6

## [ASII] Action Sequence Integrity Identification

What is the correct temporal order of the 5 narrated events? Choose the option that lists the events chronologically.

A. 1. pick up third orange -> 2. place juicer bowl -> 3. lift lid of food recycling bin -> 4. open lid of food processing machine -> 5. hold opened panel of boiler

B. 1. rotate dial of food processor -> 2. close food recycling bin's lid -> 3. remove plastic cover of scissors -> 4. move wooden stirrer -> 5. throw wooden stirrer

C. 1. pick up halves of orange -> 2. place juicer bowl -> 3. pick up milk bottle -> 4. place juicer part -> 5. throw half orange skin

D. 1. close fridge -> 2. crawl on counter top -> 3. rotate milk frother -> 4. lift lid of box of wooden stirrers -> 5. put knife

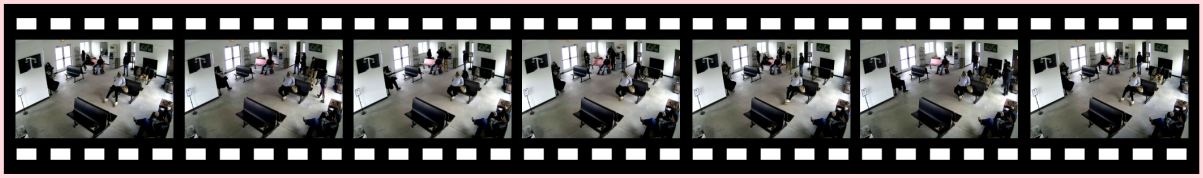
E. 1. press button -> 2. tilt milk -> 3. move milk frother -> 4. place orange -> 5. wipe cone

Figure 7

## [AGBI] Appearance-Grounded Behavior & Interactions

In the video there is exactly one individual that fits the following description. The individual appears to be male, wearing a black hoodie and khaki pants. His footwear is light-colored boots, and he is using earphones. He has short, dark hair, and there is a camouflaged backpack next to him.

What notable action does the person perform near a machine in the video?



A. Withdraws cash from an ATM.

B. Engages in a conversation with a stranger

C. Repairs a vending machine.

D. Waits and then buys a ticket from a kiosk.

E. Stops briefly near the machine and throw something into the nearby trash.

Figure 8

## [AGAR] Appearance-Grounded Attribute Recognition

In the video there is exactly one individual that fits the following description. The individual is female, wearing a long yellow scarf, a black coat, black pants, and black shoes. She has brown hair and carries a black backpack. In her hand, she is holding a white paper or booklet.

As the person walks, she passes an individual on the left wearing what kind of clothing?

A. A green jacket and brown pants

B. A black coat holding a phone

C. A red coat and white jeans

D. A yellow t-shirt and gray trousers

E. A brown sweater and shorts

Figure 9

## [AGLT] Appearance-Grounded Localization & Trajectory

In the video there is exactly one individual that fits the following description. The person appears to be male, wearing a light gray long-sleeve shirt and blue jeans. They have short dark hair and are wearing glasses. The individual is holding a phone to their ear, suggesting they are on a call.

In the video, where does the individual initially start moving from?"

A. The left edge of the frame.

B. The right edge of the frame.

C. The bottom-right quadrant.

D. The top-left corner.

E. The center of the frame.

Figure 10

## [FAM] False Object Memory

Which of the following actions did NOT occur in the video?

A. pick up knife

B. tilt food processor bowl

C. push top end of panel

D. open lid of food recycling bin

E. pick wooden stirrer

Figure 11

## [SVA] Scene Verification & Arrangement

From the correctly described shots, which is the one that appears first in the video?

A. None of the above shots happened in the video as outlined.

B. A young boy with curly brown hair in a striped shirt and a young girl with straight blonde hair in a yellow and white dress sit next to each other on a bus, with a girl in glasses visible in the background through a window with a sunlit outdoor scene.

C. A man observes a woman with long hair and a dark outfit running towards a gray car on a nighttime street with a building and trees in the background.

D. A person with brown hair wearing a white shirt and black pants sits atop a mast with a nautical flag on a yacht sailing near an island, with the coastline visible in the distance.

E. Two men in military attire, surrounded by dense foliage and smoke, are struggling together as one helps the other to stand up.

Figure 12

## [FOM] False Object Memory

Which of the following objects the camera wearer of the video did NOT interact with?

A. cupboard

B. lather

C. garlic top

D. onions

E. separated chicken piece

Figure 13

## [MEGL] Multi-Entities Grounding and Localization

The following 2 people appeared in the video:
1. The individual appears to be a male, wearing a navy blue jacket over a red shirt. He has light-colored pants and dark shoes, with a black backpack on his back. His hair is short and brown, and he often appears with his hands in his pockets or holding something like a phone.
2. The individual appears to be male, dressed in a brown hoodie and blue jeans. His hair is dark and short, and he doesn't seem to have any visible accessories. His posture is relaxed, standing with his hands in his pockets.
From which edge of the frame did they exit the scene?

A. 1 stayed in the video until the end, 2 exited through the right edge.

B. 1 exited through the left edge, 2 stayed in the video until the end.

C. 1 exited through the right edge, 2 stayed in the video until the end.

D. both exited through the left edge.

E. both stayed in the video until the end.

Figure 14

## [AC] Action Counting

How many times does the action-object pair 'open fridge door' occur in this video? Choose the correct count.

A. 3

B. 0

C. 2

D. 5

E. 1

Figure 15

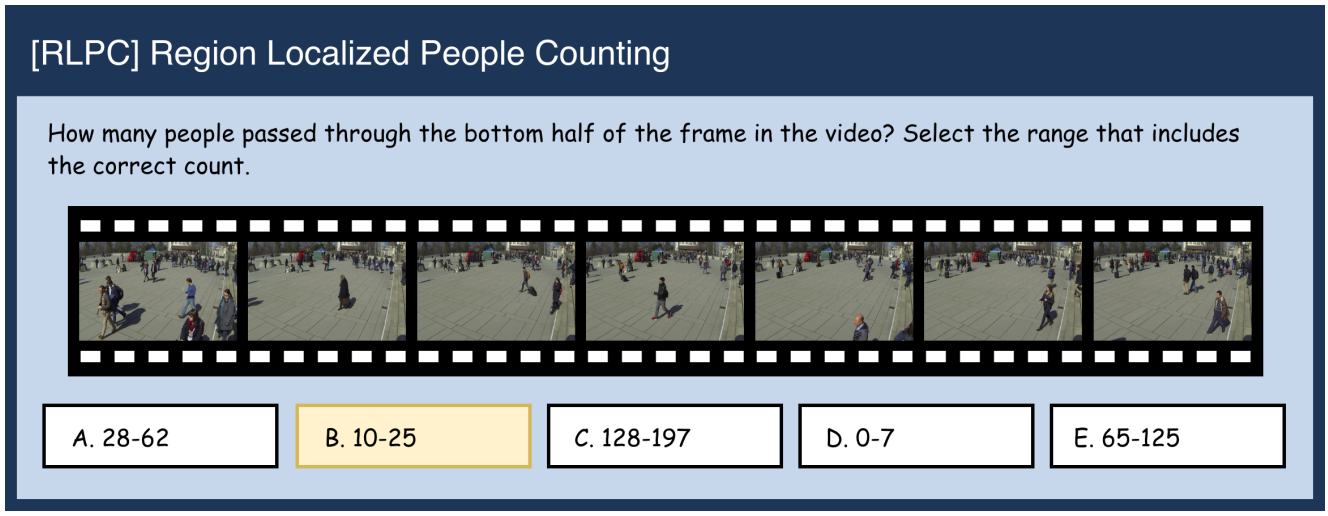

[RLPC] Region Localized People Counting
How many people passed through the bottom half of the frame in the video? Select the range that includes the correct count.
A. 28-62
B. 10-25
C. 128-197
D. 0-7
E. 65-125

Figure 16

## References


[1] Xiang An et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025. 7

[2] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 7

[3] Shuai Bai et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 5, 6, 7

[4] Tatjana Chavdarova, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Timur Bagautdinov, Louis Lettry, Pascal Fua, Luc Van Gool, and François Fleuret. Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5030–5039, 2018. 5

[5] Wei-Lin Chiang, Zhuohan Li, et al. Vicuna v1.5: An open-source chatbot, 2023. FastChat project report. 4

[6] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 7

[7] Glenn Jocher and Jing Qiu. Ultralytics yolo11, 2024. 1

[8] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 6, 7

[9] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, 2024. 8

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014. 2

[11] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision (ECCV)*, pages 38–56. Springer, 2024. 1

[12] Shuming Liu, Chen Zhao, Tianqi Xu, and Bernard Ghanem. Bolt: Boost large vision-language model without training for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3318–3327, 2025. 6, 7

[13] Shiyin Lu et al. Ovis2.5 technical report. *arXiv preprint arXiv:2508.11737*, 2025. 7

[14] Meta AI. The llama 3 herd of models, 2024. Model release report. 4

[15] OpenAI. Gpt-4o, 2024. Model card and system card. 2, 7

[16] Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Kumar Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, Jacob Chalk, Zhifan Zhu, Rhodri Guerrier, Fahd Abdelazim, Bin Zhu, Davide Moltisanti, Michael Wray, Hazel Doughty, and Dima Damen. Hd-epic: A highly-detailed egocentric video dataset. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23901–23913, 2025. 5

[17] Qwen Team. Qwen2: A family of open large language models, 2024. Alibaba Cloud. 4

[18] Qwen Team. Qwen2.5 technical report, 2024. Alibaba Cloud. 4

[19] Isaac Robinson, Peter Robicheaux, Matvei Popov, Deva Ramanan, and Neehar Peri. Rf-detr: neural architecture search for real-time detection transformers. *arXiv preprint arXiv:2511.09554*, 2025. 1, 3

[20] Bing Shuai, Alessandro Bergamo, Uta Buechler, Andrew Berneshawi, Alyssa Boden, and Joseph Tighe. Large scale real-world multi-person tracking. In *European Conference on Computer Vision*, pages 504–521. Springer, 2022. 5

[21] Tomáš Souček and Jakub Lokoč. TransNet V2: An effective deep network architecture for fast shot transition detection. *arXiv preprint arXiv:2008.04838*, 2020. 3

[22] Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. DanceTrack: Multi-object tracking in uniform appearance and diverse motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20993–21002, 2022. 1

[23] Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *CVPR*, 2025. arXiv:2502.21271. 5, 6, 7

[24] Weiyun Wang et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 7

[25] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep as-

sociation metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, 2017. 1, 3

[26] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. In *NeurIPS*, 2024. arXiv:2407.15754. 8

[27] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *CVPR*, 2021. 6, 8

[28] Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, et al. Re-thinking temporal search for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8579–8591, 2025. 7

[29] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. ByteTrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision (ECCV)*, pages 1–21. Springer, 2022. 1