

ARGOS: Who, Where, and When in Agentic Multi-Camera Person Search

Supplementary Material

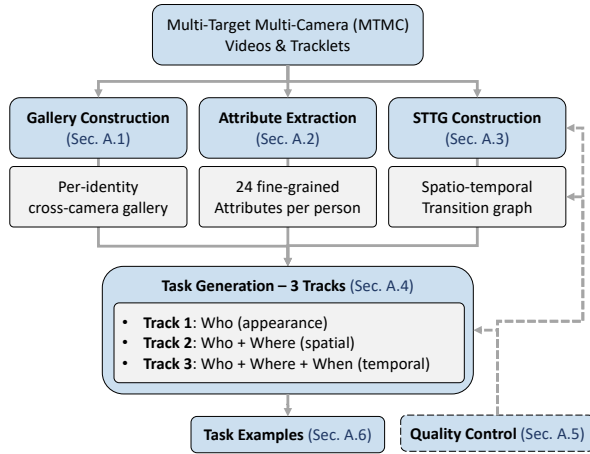


Figure 5. **ARGOS benchmark construction pipeline.** Starting from synchronized 16-camera video (MTMMC [21]), we construct per-person galleries (Sec. A.1), extract 24 visual attributes (Sec. A.2), and build a spatio-temporal transition graph (Sec. A.3). These three components feed into the task generation module (Sec. A.4), which produces 989 (Track 1), 550 (Track 2), and 1,152 (Track 3) task instances. Quality control (Sec. A.5) is applied at each stage.

ure 5 illustrates the overall pipeline, which consists of five stages followed by quality control.

- **Sec. A.1:** Gallery construction from synchronized multi-camera video.
- **Sec. A.2:** VLM-based attribute extraction with multi-model verification.
- **Sec. A.3:** Spatio-temporal transition graph (STTG) construction.
- **Sec. A.4:** Task generation for Tracks 1, 2, and 3.
- **Sec. A.5:** Quality control and human verification at each stage.
- **Sec. A.6:** Task examples and qualitative tool demonstrations for each track.

A.1. Gallery Construction from Multi-Camera Data

ARGOS constructs a per-person gallery of 10 representative images from the MTMMC dataset [21], which provides 16 synchronized cameras per environment with ground-truth person identities and per-frame bounding boxes. The gallery construction pipeline (Algorithm 1) selects images that maximize both visual quality and viewpoint diversity for downstream VLM-based attribute extraction.

Input data. The MTMMC dataset provides cross-camera identity labels as ground truth. ARGOS does not perform additional cross-camera association; instead, the pipeline focuses on selecting the most informative crops for each known identity. Raw frames span 14 scenarios across two environments (7 factory, 7 university), each with 16 synchronized cameras.

Five-stage pipeline. Algorithm 1 describes the full procedure. Stage 1 filters persons appearing in fewer than 3 unique cameras, ensuring multi-view coverage. Stage 2 samples frames from the temporally safe interior of each tracklet (excluding the first and last 5%), avoiding truncated appearances at tracklet boundaries. Stage 3 applies a three-tier crop quality filter: (i) basic geometry checks (minimum size, aspect ratio), (ii) a pose estimation gate using YOLOv8-Pose [19] requiring at least 10 of 17 COCO keypoints with confidence ≥ 0.7 , and (iii) pose-aware target verification that confirms the detected person is centered ($|c_x - \text{center}| \leq 0.15W$) and occupies an appropriate image fraction ($0.40 \leq \text{coverage} \leq 1.05$). The pose gate directly verifies that the person’s body is sufficiently visible for VLM attribute extraction, going beyond simple resolution thresholds.

388 Supplementary Material

389 This supplementary document provides additional details
390 referenced in the main paper. It is organized into five
391 parts:

- 392 • **Part A** (Sec. A): Benchmark construction pipeline,
393 including gallery construction, attribute extraction,
394 STTG construction, task generation, quality control,
395 and track examples with qualitative tool demonstra-
396 tions.
- 397 • **Part B** (Sec. B): Benchmark overview details, includ-
398 ing the three-track overview figure and benchmark
399 statistics table omitted from the main paper for space.
- 400 • **Part C** (Sec. C): Agent architecture details, including
401 the full system architecture, information boundary, and
402 tool registry.
- 403 • **Part D** (Sec. D): Additional experimental results, in-
404 cluding parsing analysis, difficulty analysis, failure
405 case analysis, and behavioral statistics.
- 406 • **Part E** (Sec. E): Additional benchmark details, in-
407 cluding the full attribute taxonomy, witness simulator
408 templates, and environment STTG details.

409 A. Benchmark Construction Pipeline

410 This section details the end-to-end construction pipeline
411 of the ARGOS benchmark, from raw multi-camera data
412 to the final 2,691 task instances across three tracks. Fig-

Algorithm 1 Gallery Construction Pipeline

Require: GT annotations (per-camera, per-frame bounding boxes), raw video frames (16 cameras \times 14 scenarios)

Ensure: Per-person gallery (10 images, 960×512 grid)

1: **Params:** CAM_MIN=3, SAFE_ZONE=0.05, FRAMES_SAMPLE=30, MIN_KP=10, KP_CONF=0.7, CENTER_THRESH=0.15, COV_MIN=0.40, COV_MAX=1.05, AREA_CUT=0.5, BEST_K=3, TOTAL_K=10

2: **for** each person p in GT **do**
 3: **Stage 1:** if $|\text{unique_cameras}(p)| < \text{CAM_MIN}$:
 reject p
 4: **for** each camera c observing p **do**
 5: **Stage 2:** $\text{safe} \leftarrow \text{frames}[\lceil 0.05 \cdot |\text{frames}| \rceil : \lfloor 0.95 \cdot |\text{frames}| \rfloor]$
 6: $\text{sampled} \leftarrow \text{uniform_sample}(\text{safe}, 30)$
 7: **for** each frame f in sampled **do**
 8: **Stage 3a:** Check geometry (margin, size, aspect ratio)
 9: **Stage 3b:** $\text{kps} \leftarrow \text{YOLOv8-Pose}(f)$; reject if $|\text{valid_kps}| < 10$
 10: **Stage 3c:** Reject if center offset $> 0.15W$ or coverage $\notin [0.40, 1.05]$
 11: **end for**
 12: **end for**
 13: **Stage 4:** $\text{pool} \leftarrow \text{sort by area, filter } \geq 0.5 \times \max$
 14: Select top 3 from best camera (anchor) + 7 via round-robin (diversity)
 15: **Stage 5:** Letterbox to 192×256 ; assemble 5×2 grid (960×512)
 16: **end for**

459 Stage 4 selects the final 10 images through a quality-
 460 diversity strategy: the top 3 images from the highest-
 461 quality camera serve as anchor frames (Phase 1), fol-
 462 lowed by 7 images sampled in round-robin order across
 463 remaining cameras (Phase 2). This ensures at least one
 464 high-resolution frontal/side view alongside diverse view-
 465 points. Stage 5 applies letterboxing (192×256 per crop)
 466 and assembles a 5×2 grid (960×512), which becomes
 467 the input to the attribute extraction module.

468 **Gallery statistics.** Table 4 reports per-scenario gallery
 469 sizes. The pipeline yields 1,273 person galleries across
 470 14 scenarios, with an average of 5.9 cameras per per-
 471 son (range: 4.9–7.4). The dominant rejection mode is
 472 *fail_no_person* (pose model detects no valid per-
 473 son), confirming that the quality gates correctly discard
 474 severely occluded or low-quality crops.

Table 4. Per-scenario gallery statistics. All persons have ≥ 3 camera views and 10 selected images.

Scenario	# IDs	Env.	Scenario	# IDs	Env.
s01	90	Factory	s34	95	Univ.
s10	94	Factory	s35	93	Univ.
s11	87	Factory	s36	87	Univ.
s13	90	Factory	s38	97	Univ.
s16	90	Factory	s39	101	Univ.
s18	73	Factory	s42	97	Univ.
s20	84	Factory	s47	95	Univ.
Factory total: 608			University total: 665		
Grand total: 1,273 persons			Avg. cameras/person: 5.9		

A.2. Attribute Extraction Protocol

We extract 24 visual attributes per person using a three-
 stage VLM pipeline (Algorithm 2). The pipeline pro-
 cesses each person’s 10-frame gallery grid and produces
 a structured JSON annotation with reasoning traces.

Stage 1: Multi-model parallel extraction. Six VLMs
 (GPT-5.2, GPT-5, GPT-5-mini, GPT-4o, GPT-4.1, GPT-
 4o-mini) independently extract 24 attributes from the
 same gallery grid image. Each model receives a struc-
 tured prompt (v7, refined over seven iterations) consist-
 ing of four phases: (i) context injection with scenario-specific
 intelligence (e.g., shadow-induced color darkening for
 outdoor factory cameras) and per-camera environment
 labels; (ii) target isolation via consensus building across
 the 10 frames, filtering occluders and distractors; (iii) an-
 chor selection, choosing the highest-quality frontal/side
 views as reference frames, with illumination normaliza-
 tion across different cameras; and (iv) head-to-toe at-
 tribute extraction following a *fact-first* paradigm where
 visual observations are recorded before classification,
 preventing premature label assignment.

A key design principle is the separation of attribute
 extraction logic: color and garment attributes are de-
 termined by *majority vote* across anchor frames, while
 accessories (backpack, phone, hat) use *existential proof*,
 where presence in any single frame suffices. This distinc-
 tion reflects the different evidence requirements: stable
 attributes need cross-view consistency, while intermit-
 tently visible items only need one clear observation.

Stage 2: Uncertainty-driven self-verification. For
 each attribute, we check agreement across the six models.
 When all models agree, the consensus value is accepted.
 When disagreement exists, a frontier model (GPT-5.2
 Pro) re-evaluates the attribute by inspecting the gallery
 grid alongside the conflicting predictions. This second-
 pass verification resolves most ambiguities without hu-
 man intervention.

Algorithm 2 VLM-based Attribute Extraction Pipeline

Require: Gallery grid image I (960×512), scenario context

Ensure: 24 attributes per person (structured JSON)

```

1:  $\mathcal{M} \leftarrow \{\text{GPT-5.2, GPT-5, GPT-5-mini, GPT-4o, GPT-4.1, GPT-4o-mini}\}$ 
2:  $\text{prompt} \leftarrow \text{BUILDPROMPT}(\text{scenario, camera\_specs, } I) \{v7 \text{ template}\}$ 
3: for each model  $m \in \mathcal{M}$  do
4:   {Stage 1: Parallel extraction}
5:    $\text{result}[m] \leftarrow m.\text{extract}(\text{prompt})$ 
6:   // Internal: consensus  $\rightarrow$  anchor selection  $\rightarrow$  fact-first extraction
7: end for
8: for each attribute  $a$  in schema do
9:   {Stage 2: Self-verification}
10:   $V \leftarrow \{\text{result}[m][a] : m \in \mathcal{M}\}$ 
11:  if all values in  $V$  agree then
12:     $\text{final}[a] \leftarrow \text{majority}(V)$ 
13:  else
14:     $\text{final}[a] \leftarrow \text{GPT-5.2-Pro.verify}(I, a, V)$ 
15:  end if
16: end for
17: Stage 3: Expert review for remaining uncertain attributes

```

512 **Stage 3: Expert review.** Attributes that remain uncertain after Stage 2 are reviewed by research team members
513 with computer vision expertise, who compare the gallery
514 grid image against predicted values. The final annotations
515 achieve 97.8% inter-model consistency, with only
516 1.17% of all attribute values (343 out of 29,280 total
517 annotations) marked as Uncertain.
518

519 **Prompt evolution.** The extraction prompt was refined
520 over seven versions. Iterative improvements included
521 introducing the reasoning trace (<think> block), ex-
522 plicit data construction rules, scenario intelligence injec-
523 tion, camera knowledge base, and the fact-first extraction
524 paradigm. The cumulative effect was a 42% reduction in
525 attribute mismatch rate relative to the initial version.

526 A.3. STTG Construction Details

527 The spatio-temporal transition graph (STTG) captures
528 how people move between cameras over time. We con-
529 struct the STTG through a three-phase pipeline: transi-
530 tion extraction from tracking ground truth, priority-based
531 labeling with human-in-the-loop curation, and directed-
532 edge aggregation with statistical characterization.

533 **Phase 1: Transition extraction.** Starting from the
534 MTMMC ground-truth tracking annotations, we extract
535 person-level trajectories as sequences of camera visits.

Table 5. Per-scenario trajectory statistics. Avg. cameras/person across all scenarios: 5.9.

Scenario	IDs	Rows	Avg Seq Len	Max Seq	Avg Cams
s01	117	1,339	11.4	20	7.0
s10	128	1,427	11.1	22	6.8
s11	121	1,323	10.9	21	6.4
s13	136	1,272	9.4	21	5.8
s16	118	1,446	12.3	21	7.4
s18	122	1,041	8.5	19	5.7
s20	124	1,256	10.1	22	6.1
s34	187	1,536	8.2	17	5.3
s35	152	1,162	7.6	20	5.1
s36	164	1,321	8.1	20	5.2
s38	159	1,169	7.4	18	4.7
s39	158	1,367	8.7	18	5.8
s42	154	1,421	9.2	20	5.6
s47	157	1,300	8.3	20	5.2

536 *A frame gap heuristic* (threshold: 100 frames \approx 4.3s
537 at 23 FPS) splits long tracking sequences at temporal
538 gaps to distinguish genuine re-entries from occlusion re-
539 recovery, ensuring physical continuity within each visit
540 segment. For each pair of consecutive camera visits by
541 the same person, we compute the raw transit time (entry
542 time at the next camera minus exit time at the current
543 camera) and preserve full floating-point precision to sup-
544 port sub-second overlap analysis. This yields 19,321
545 raw transitions across all 14 scenarios from 1,997 unique
546 person-scenario pairs. Table 5 reports per-scenario tra-
547 jectory statistics.

548 **Phase 2: Priority-based labeling.** Each transition is
549 classified into TRUST, WARN, or FAIL through Algo-
550 rithm 3, a seven-level priority hierarchy that handles
551 ambiguous cases systematically. Two inputs require expert
552 curation: (i) a manual error list identifying tracking
553 annotation errors (e.g., identity switches where the same
554 person appears at two distant cameras simultaneously),
555 compiled by inspecting flagged trajectories; and (ii) a
556 suspicious simultaneous presence set, where the same
557 person is recorded at multiple cameras at the same time.

558 The three edge types encode the physical relationship
559 between camera pairs: **OVERLAP** pairs share a field
560 of view (near-zero or negative transit times are normal);
561 **SOFT_ADJ** pairs are physically adjacent (separated by
562 a door or wall) without visual overlap; **TRAVEL** pairs
563 require walking between distant locations. **OVERLAP**
564 pairs are pre-defined based on camera placement (fac-
565 tory: 8 pairs, university: 11 pairs) and confirmed through
566 Union-Find clustering. **SOFT_ADJ** pairs are identified by
567 the research team based on physical camera layout (fac-
568 tory: 2 pairs, university: 2 pairs), with acceptance thresh-
569 olds computed dynamically from the 95th percentile of
570 negative transit-time distributions, capped at 2 seconds.

Algorithm 3 Transition Labeling Pipeline (Priority-based)

Require: Transition T , Manual error set M , Simultaneous set S , Overlap pairs O_{hard} , Soft-adjacency pairs O_{soft} , Dynamic threshold map Θ_{adj}

Ensure: (status $\in \{\text{TRUST, WARN, FAIL}\}$, reason)

```

1: Params:  $\varepsilon=0.1\text{s}$ ,  $\tau_{\text{gap}}=0.3\text{s}$ ,  $\tau_{\text{dur}}=0.5\text{s}$ 
2: if  $T.\text{id} \in M$  then
3:   return (FAIL, ManualError)
4: end if
5: if  $T.\text{tracklet} \in S$  then
6:   return (FAIL, Simultaneous)
7: end if
8: if  $T.\text{edge} \in O_{\text{hard}}$  then
9:   return (TRUST, Overlap)
10: end if
11: if  $T.\text{edge} \in O_{\text{soft}}$  then
12:    $\theta \leftarrow \Theta_{\text{adj}}[T.\text{edge}]$  {P95-based, capped at 2s}
13:   if  $T.\text{transit} < -\varepsilon$  then
14:     if  $|T.\text{transit}| \leq \theta$  then
15:       return (WARN, SoftAdjNeg)
16:     else
17:       return (FAIL, BeyondAdj)
18:     end if
19:   end if
20: end if
21: if  $T.\text{transit} < -\varepsilon$  then
22:   return (FAIL, ResidualNeg)
23: end if
24: if  $0 \leq T.\text{transit} < \tau_{\text{gap}}$  then
25:   return (WARN, TinyGap)
26: end if
27: if  $T.\text{duration} \leq \tau_{\text{dur}}$  then
28:   return (WARN, ShortSeq)
29: end if
30: return (TRUST, Default)

```

571 The resulting classification is: TRUST 18,151 transi-
572 tions (94%), WARN 229 (1%), FAIL 941 (5%). Only
573 TRUST transitions are used for downstream graph con-
574 struction.

575 **Phase 3: Directed-edge aggregation.** For each di-
576 rected camera pair ($c_i \rightarrow c_j$) with TRUST transitions,
577 we compute transition-time statistics: minimum, median,
578 maximum, mean, and standard deviation. The constraint
579 $t_{\text{min}} \leq t_{\text{med}} \leq t_{\text{max}}$ is verified for all edges. Table 6 sum-
580 marizes the edge-type distribution. The factory STTG
581 contains 110 edges and the university STTG contains 149
582 edges, consistent with the values reported in the main
583 paper (Sec. 2). Among non-self-loop edges, 31 (factory)
584 and 40 (university) have ≥ 20 observations, providing sta-
585 tistically reliable transition-time estimates for temporal

Table 6. STTG edge-type distribution (incl. 16 self-loops per environment).

Environment	OVERLAP	SOFT_ADJ	TRAVEL	Self-loops	Total
Factory	16	4	74	16	110
University	22	4	107	16	149

reasoning in Track 3.

Zone definitions. Cameras connected by OVERLAP edges are grouped into zones via Union-Find. Adjacent zones linked by SOFT_ADJ edges form composite zones for Track 2 spatial disambiguation. The factory environment has 9 zones (3 composite zones), and the university has 6 zones (2 composite zones). Each zone receives a natural-language label describing its physical location (e.g., “the warehouse area,” “the outdoor plaza”), assigned by the research team based on camera placement. These labels are used in Track 2 disambiguation questions.

University STTG. The main paper (Fig. 2) shows the factory STTG. The university campus STTG (149 edges, 16 cameras, 6 zones) is presented in Sec. E.3 with zone definitions and structural comparison to the factory.

A.4. Task Generation

Task generation follows a two-stage architecture shared across all three tracks. **Stage 1** (deterministic) computes the ground-truth disambiguation path for each target person using the gallery attribute database, trajectories, and STTG. **Stage 2** (LLM-based) wraps the structured simulation log into natural language dialogue. This separation ensures that benchmark correctness does not depend on LLM behavior; Stage 2 only transforms surface form while preserving ground-truth semantics.

A.4.1. Track 1: Information-Gain Disambiguation

Track 1 evaluates an agent’s ability to identify a person through sequential attribute questions. Algorithm 4 describes the full generation procedure.

Saliency scoring. For each scenario, we compute TF-IDF saliency scores over attribute values:

$$\text{Saliency}(a, v) = \frac{1}{\text{freq}(v)} \times \log \frac{N}{\text{valid_count}(a)}, \quad (2)$$

where N is the gallery size and $\text{valid_count}(a)$ excludes Uncertain and None values. Rare attribute values receive higher saliency, guiding clue selection toward discriminative starting points.

Clue selection. Each target is assigned to a difficulty mode via a deterministic per-target seed ($\text{id} \times 1337 + 42$):

625 *Hard mode* (35% of targets) selects common, low-
626 saliency attributes to create large initial candidate pools
627 (30–90 persons); *Balanced mode* (65%) selects salient
628 attributes for smaller pools (5–30 persons). Ten at-
629 tributes with low discriminative utility as initial clues
630 (e.g., shoes_color, age_style) are excluded from
631 clue selection but remain available for IG-based question-
632 ing.

633 **Greedy information gain simulation.** At each turn,
634 the algorithm selects the attribute that maximizes a pe-
635 nalized entropy score over the current candidate set \mathcal{C}_t :

$$636 \quad IG(a) = H_{\text{value}}(a | \mathcal{C}_t) \times (1 - \alpha \cdot p_{\text{uncertain}}(a)), \quad (3)$$

637 where $H_{\text{value}}(a | \mathcal{C}_t) = -\sum_v p(v) \log_2 p(v)$ is the value
638 distribution entropy of attribute a across current candi-
639 dates, and $\alpha = 0.5$ penalizes attributes with high Un-
640 certain rates. Maximizing this score selects the attribute
641 whose values most evenly partition the candidate set,
642 yielding the greatest expected reduction per question.
643 The simulation applies the target’s ground-truth attribute
644 value at each turn and filters candidates accordingly. Of
645 the 24 gallery attributes, 21 are used in the IG simula-
646 tion; three (hair_visibility, leg_visibility,
647 body_features) are excluded due to extreme class
648 imbalance, which provides negligible discrimination. All
649 24 attributes remain in the gallery database accessible to
650 agents during evaluation.

651 **Uniqueness guarantee.** Every generated task satis-
652 fies $|\mathcal{C}_{\text{final}}| = 1$ and $\mathcal{C}_{\text{final}} = \{g^*\}$, where g^* is
653 the target person. This is verified programmatically:
654 across all 14 scenarios, zero tasks fail this check
655 (ambiguous_target = 0). Tasks where the candidate
656 pool falls below 5 after clue application are discarded
657 (too_few_candidates: 284 cases), ensuring a mini-
658 mum level of disambiguation challenge.

659 **Difficulty labeling.** Difficulty scores combine three
660 factors:

$$661 \quad \text{base} = 0.45 \cdot \frac{|\mathcal{C}_0|}{90} + 0.40 \cdot \frac{T}{10} + 0.15 \cdot r_u, \quad (4)$$

662 where $|\mathcal{C}_0|$ is the initial candidate count, T is the number
663 of IG turns, and r_u is the uncertain attribute ratio. A turn
664 gating factor ($\times 0.7$ if $T < 3$) discourages labeling short
665 paths as Hard. The final score is discretized into Easy,
666 Medium, and Hard using global percentile thresholds
667 (33rd and 67th).

668 **Natural language wrapping (Stage 2).** GPT-4o (tem-
669 perature 0.3) converts each structured simulation log into
670 a natural witness–agent dialogue. The model is instructed

Algorithm 4 Track 1 — Information-Gain Disambigua-
tion

Require: Gallery DB G (24 attrs \times N persons), seed
 $= 42$, $\alpha = 0.5$

Ensure: Task set \mathcal{T}_1 with natural language dialogues

- 1: saliency \leftarrow TF-IDF(G)
- 2: **for** each target $g^* \in G$ **do**
- 3: mode \leftarrow Hard (35%) or Balanced (65%) {per-
target seed}
- 4: clue \leftarrow SELECTCLUE(g^* , saliency, mode)
- 5: $\mathcal{C}_0 \leftarrow \{g \in G : g \text{ matches clue}\}$
- 6: **if** $|\mathcal{C}_0| \notin [5, 90]$ **then**
- 7: skip
- 8: **end if**
- 9: **for** $t = 1$ to 10 **do**
- 10: {Greedy IG simulation}
- 11: **if** $|\mathcal{C}_t| \leq 1$ **then**
- 12: break
- 13: **end if**
- 14: $a^* \leftarrow \arg \max_a H_{\text{value}}(a | \mathcal{C}_t) \cdot (1 - 0.5 \cdot$
 $p_{\text{unc}}(a))$
- 15: $\mathcal{C}_{t+1} \leftarrow \{g \in \mathcal{C}_t : g[a^*] = g^*[a^*]\}$
- 16: **end for**
- 17: **if** $|\mathcal{C}_{\text{final}}| = 1 \wedge \mathcal{C}_{\text{final}} = \{g^*\}$ **then**
- 18: accept
- 19: **end if**
- 20: **end for**
- 21: **Stage 2:** GPT-4o wraps each simulation log into NL
dialogue

Table 7. Track 1 turn distribution and difficulty breakdown (989 tasks total).

	Turns					Difficulty		
	1	2	3	4	5	Easy	Med	Hard
Count	148	472	285	80	4	340	326	323
%	15.0	47.7	28.8	8.1	0.4	34.4	33.0	32.7

to produce full natural sentences with no raw attribute
labels. Quality metrics confirm zero robotic exact-match
responses (REMR = 0.0%), perfect ground-truth keyword
preservation (GKPR = 100.0%), and a mean witness ut-
terance length of 7.2 words.

Track 1 statistics. Table 7 reports the turn distribution.
The average oracle path length is 2.31 turns, with 47.7%
of tasks requiring exactly 2 turns. The average search-
space reduction rate (SSRR) is 69.7% per turn.

A.4.2. Track 2: Zone-based Spatial Disambiguation

Track 2 adds spatial reasoning to the attribute-based dis-
ambiguation of Track 1. Each task is anchored to a tar-
get person observed at a specific camera within a multi-
camera zone, and the agent must combine appearance
questions with location questions to identify the target.

686 Algorithm 5 describes the generation procedure.

687 **Zone-based candidate selection.** For each target, the
688 pipeline identifies the primary camera (longest stay dura-
689 tion) and its containing zone. Targets in singleton zones
690 (a single camera) are excluded, as spatial disambiguation
691 is impossible. Candidates are persons observed in the
692 same zone within a ± 300 -second window of the target’s
693 presence.

694 **Disambiguation path construction.** At each turn, the
695 algorithm considers both spatial and attribute questions.
696 A spatial question uses a pre-defined disambiguation tree
697 for the target’s zone, partitioning cameras into sub-areas
698 based on physical layout (e.g., “Was it deep inside the
699 warehouse, or near the entrance?”). An attribute question
700 is selected by the same IG-based procedure as Track 1.
701 The algorithm chooses spatial over attribute when the
702 spatial elimination count is at least half the attribute elim-
703 ination count ($\text{spatial_elim} \geq 0.5 \times \text{attr_elim}$), giving
704 priority to spatial reasoning while maintaining efficiency.

705 **Quality filter.** Tasks with zero spatial turns are dis-
706 carded, ensuring every Track 2 task genuinely requires
707 spatial reasoning and is not solvable by attributes alone.
708 This filter removes 602 of 1,218 candidate tasks (49.4%),
709 yielding 550 accepted tasks.

710 **Natural language wrapping (Stage 2).** GPT-5.2 con-
711 verts the structured path into natural dialogue, with strict
712 rules: no camera IDs, no zone IDs, and no robotic lan-
713 guage. Location descriptions use natural language (e.g.,
714 “near the tall shelves in the back”) rather than technical
715 identifiers.

716 **Track 2 statistics.** The 550 accepted tasks average
717 1.14 spatial turns and 0.91 attribute turns (2.05 total or-
718 acle turns). Difficulty distribution: Easy 149 (27.1%),
719 Medium 282 (51.3%), Hard 119 (21.6%). Difficulty is de-
720 termined by initial candidate count and total turns: Easy
721 ($|\mathcal{C}_0| \leq 5$, turns ≤ 2), Medium ($|\mathcal{C}_0| \leq 10$, turns ≤ 4),
722 and Hard (otherwise).

723 A.4.3. Track 3: STTG-based Temporal Feasibility

724 Track 3 evaluates an agent’s ability to reason about tem-
725 poral plausibility using the STTG. Each task presents
726 two sightings of the target at different cameras, and the
727 agent must determine which candidates could have made
728 the observed transition. Algorithm 6 describes the full
729 procedure.

730 **STTG-Only Trust principle.** A fundamental design
731 constraint is that temporal reasoning uses only directly

Algorithm 5 Track 2 — Zone-based Spatial Disambiguation

Require: Gallery G , Trajectories Π , Zone definitions Z

Ensure: Task set \mathcal{T}_2 (spatial_turns ≥ 1)

```

1: for each target  $g^* \in G$  do
2:    $c^* \leftarrow \text{primary\_camera}(\Pi[g^*]); \quad z^* \leftarrow \text{zone}(c^*)$ 
3:   if  $|z^*.cameras| < 2$  then
4:     skip
5:   end if
6:   candidates  $\leftarrow$  persons in  $z^*$  within  $\pm 300$ s of  $g^*$ 
7:   clue  $\leftarrow \text{SELECTCLUE}(g^*, \text{candidates}); \quad \mathcal{C}_0 \leftarrow$ 
   filter by clue
8:   for  $t = 1$  to 6 do
9:     if  $|\mathcal{C}_t| \leq 1$  then
10:      break
11:    end if
12:    Compute spatial_elim (zone tree) and attr_elim (IG)
13:    if spatial possible and spatial_elim  $\geq 0.5 \times$ 
   attr_elim then
14:      Apply spatial question
15:    else
16:      Apply attribute question (best IG)
17:    end if
18:  end for
19:  if resolved to  $\{g^*\}$  and spatial_turns  $\geq 1$  then
20:    accept
21:  end if
22: end for
23: Stage 2: GPT-5.2 wraps path into NL (no cam-
   era/zone IDs)

```

observed STTG edges; multi-hop path inference is ex- 732
cluded. If no direct edge exists between a camera pair, 733
the candidate’s feasibility is classified as UNKNOWN 734
and excluded from the ground truth. This ensures that all 735
temporal judgments are grounded in empirical transition- 736
time statistics. 737

Two-sighting selection. For each target, the pipeline 738
finds the best pair of consecutive camera visits from the 739
target’s trajectory. Candidate transitions must have a 740
TRAVEL or SOFT_ADJ edge with ≥ 20 observations and 741
a positive transit time. A scoring function prioritizes 742
TRAVEL edges (score +10 vs. +3 for SOFT_ADJ), mod- 743
erate median times (10–60s, +5), high sample counts 744
(+4), and actual times within the statistical range (+3). 745

Temporal feasibility classification. For each candi- 746
date in the initial pool, a four-stage check determines 747
feasibility: 748

1. **Presence:** Was the candidate observed at both cam- 749

- 750 eras? If not: IMPOSSIBLE (NOT_PRESENT).
 751 2. **Time ordering:** Is the candidate’s inter-camera gap
 752 $< -5s$? If so: IMPOSSIBLE (TIME_REVERSAL).
 753 3. **Edge existence:** Does a direct STTG edge exist? If
 754 not: UNKNOWN (excluded).
 755 4. **Plausibility (margin = 2.0):** Is the gap too fast
 756 ($< t_{\min}/2.0$) or too slow ($> t_{\max} \times 2.0$)? If so:
 757 IMPOSSIBLE. Otherwise: FEASIBLE.

758 **Quality filter.** Tasks must contain at least one
 759 candidate eliminated by genuine temporal reasoning
 760 (TIME_REVERSAL or TOO_SLOW), excluding trivial
 761 presence-based filtering. After applying the quality fil-
 762 ter, 1,152 of 1,218 candidate tasks pass (94.6%). Across
 763 accepted tasks, the average number of temporally impos-
 764 sible candidates is 4.0 per task, with TIME_REVERSAL
 765 accounting for 75% and TOO_SLOW for 25%. No
 766 TOO_FAST eliminations occur in the data, consistent
 767 with the absence of physically implausible high-speed
 768 movements.

769 **Time vaguification.** To prevent agents from exploiting
 770 exact timestamps, the dialogue wrapper converts precise
 771 transition times to vague expressions: 0–30s \rightarrow “almost
 772 at the same time,” 30–120s \rightarrow “a moment later,” 120–
 773 300s \rightarrow “a few minutes later,” and so on.

774 **Track 3 statistics.** The 1,152 tasks have an average or-
 775 acle path of $\tau^* = 1.89$ turns (1 temporal + 0.89 attribute).
 776 Difficulty distribution: Easy 145 (12.6%), Medium 702
 777 (60.9%), Hard 305 (26.5%). The skew toward Medium
 778 and Hard reflects the inherent complexity of temporal
 779 reasoning: Easy requires both few attribute candidates
 780 (≤ 6) and multiple temporal eliminations (≥ 2).

781 **Summary.** Consolidated benchmark statistics across
 782 all three tracks are provided in Table 9 (Part B).

783 A.5. Quality Control and Human Verification

784 Each construction stage employs verification methods
 785 appropriate to its nature: automated quality gates for
 786 deterministic stages, multi-model ensemble agreement
 787 for VLM-dependent stages, and algorithmic guarantees
 788 for task generation. Two stages involve expert review
 789 supported by dedicated inspection tools (Figures 6 and 7).
 790 Table 8 summarizes the quality control at each stage.

791 **Gallery construction (A.1).** The five-stage pipeline is
 792 fully automated. Quality is ensured by the pose-based
 793 gates described in Sec. A.1. Rejection statistics con-
 794 firm that the dominant failure mode is the absence of a
 795 detectable person (`fail.no.person`), indicating that

Algorithm 6 Track 3 — STTG-based Temporal Feasibil-
 ity

Require: Gallery G , Trajectories Π , STTG \mathcal{T}

Ensure: Task set \mathcal{T}_3 (true temporal impossible ≥ 1)

```

1: Params: MIN_EDGE_COUNT = 20, MARGIN
   = 2.0
2: for each target  $g^* \in G$  do
3:   Find best two-sighting pair  $(c_1, c_2, t_1, t_2)$  from
    $\Pi[g^*]$ 
4:    $clue \leftarrow \text{SELECTATTRS}(g^*)$ ;  $\mathcal{C}_0 \leftarrow$  filter by clue

5:   for each candidate  $c \in \mathcal{C}_0$  do
6:     if  $c$  not at  $c_1$  or  $c_2$  then
7:       IMPOSSIBLE (NOT_PRESENT)
8:     end if
9:      $\delta \leftarrow c.\text{enter}(c_2) - c.\text{exit}(c_1)$ 
10:    if  $\delta < -5s$  then
11:      IMPOSSIBLE (TIME_REVERSAL)
12:    end if
13:     $e \leftarrow \mathcal{T}.\text{lookup}(c_1 \rightarrow c_2)$ 
14:    if  $e$  is None then
15:      UNKNOWN
16:    end if
17:    if  $0 < \delta < e.t_{\min}/2.0$  then
18:      IMPOSSIBLE (TOO_FAST)
19:    else if  $\delta > e.t_{\max} \times 2.0$  then
20:      IMPOSSIBLE (TOO_SLOW)
21:    else
22:      FEASIBLE
23:    end if
24:  end for
25:  Require  $\geq 1$  TIME_REVERSAL or TOO_SLOW
   candidate
26:  Disambiguate feasible candidates via attribute IG
27:  if resolved to  $\{g^*\}$  then
28:    accept
29:  end if
30: end for
31: Stage 2: GPT-5.2 wraps path into NL (vague time,
   no camera IDs)

```

the filters correctly discard low-quality crops rather than
 valid person images.

Attribute extraction (A.2). Quality relies on the three-
 stage pipeline in Sec. A.2. The six-model ensemble
 provides implicit inter-annotator agreement: 97.8% of
 attributes achieve unanimous consensus across all six
 VLMs. The remaining 1.17% of attribute annotations
 (343 out of 29,280 total) are flagged as uncertain and
 reviewed by research team members through the VLM
 Attribute Inspector (Figure 6). This tool displays the 10-
 frame gallery grid alongside attribute predictions from all

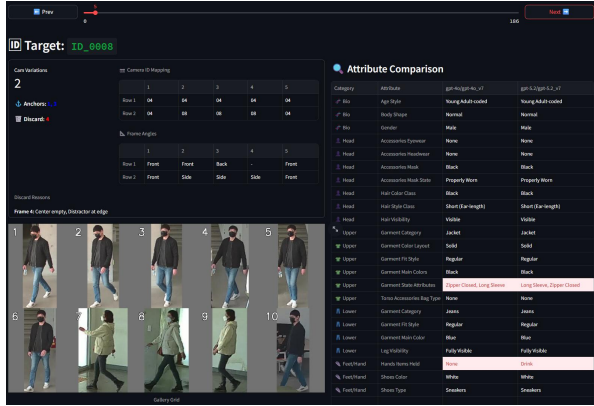
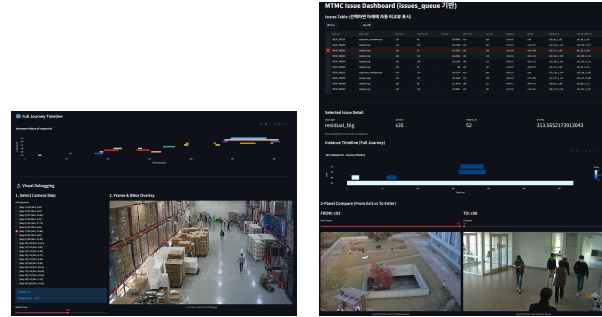


Figure 6. **VLM Attribute Inspector** used for expert review of attribute extraction. **Left:** 10-frame gallery grid for a single person (5×2, 960×512). **Right:** Attribute predictions from six VLMs displayed side-by-side. Red cells indicate inter-model disagreement; yellow cells indicate uncertain values. The research team reviewed all flagged instances through this interface to produce the final attribute annotations (97.8% inter-model consistency).



(a) Trajectory Inspector (b) STTG Issue Dashboard

Figure 7. **STTG verification tools**. (a) Trajectory Inspector: a Gantt chart shows one person’s movement across cameras over time (top); selecting a specific camera and frame displays the original video frame with GT bounding-box overlay (bottom), confirming the person’s identity at that location. (b) STTG Issue Dashboard: suspicious transitions are ranked by severity (top); a side-by-side view compares the exit frame from one camera with the enter frame at the next camera, both with bounding-box overlay (bottom), to verify whether the depicted person is the same individual.

807 six models, with automatic highlighting of inter-model
808 disagreements (red) and uncertain values (yellow). An-
809 notators visually compare each flagged attribute against
810 the gallery images and, when needed, inspect model rea-
811 soning traces to resolve conflicts.

812 **STTG construction (A.3).** The priority-based labeling
813 pipeline (Algorithm 3) incorporates two human-curated
814 inputs: a manual error list (tracking annotation errors
815 identified by trajectory inspection) and a suspicious si-
816 multaneous presence set. Expert curation is supported
817 by two inspection tools (Figure 7). The *Trajectory In-*
818 *spector* (Figure 7a) displays each person’s full movement
819 as a Gantt chart across cameras over time, with frame-
820 level visual confirmation via bounding-box overlay on
821 the original video frames. This tool is used to compile
822 the manual error list by identifying tracking annotation
823 errors (e.g., identity switches where the same ID ap-
824 pears at two distant cameras simultaneously). The *STTG Issue*
825 *Dashboard* (Figure 7b) ranks all suspicious transitions
826 by severity and provides a side-by-side frame com-
827 parison (exit frame from camera A vs. enter frame at cam-
828 era B) with bounding-box overlay, enabling experts to
829 verify whether the same person is depicted in both frames.
830 These tools were used to curate the 19,321 raw transi-
831 tions, yielding a 94% TRUST rate (18,151 transitions)
832 with 5% classified as FAIL (941 transitions). SOFT_ADJ
833 camera pairs (factory: 2 pairs, university: 2 pairs) and
834 their dynamic thresholds are specified by the research
835 team based on physical camera placement.

Task generation (A.4). Stage 1 is fully deterministic 836
with no human judgment. Structural guarantees include: 837
Track 1 uniqueness (0 failures across 989 tasks), Track 2 838
spatial turn requirement (100%), and Track 3 genuine 839
temporal elimination requirement (100%). Stage 2 dia- 840
logue quality is verified through automated metrics and 841
manual spot-checks. For Track 1, we measure Robotic 842
Exact-Match Rate (REMR = 0.0%), GT Keyword Preser- 843
vation Rate (GKPR = 100.0%), and average witness ut- 844
terance length (7.2 words). For Tracks 2 and 3, we verify 845
zero camera-ID or zone-ID leakage in all generated dia- 846
logues, confirm the absence of robotic language patterns, 847
and spot-check approximately 50 dialogues per track for 848
naturalness and GT fidelity. 849

A.6. Track Examples and Qualitative Tool 850
Demonstrations 851

We present one representative task instance per track, 852
showing the full agent interaction trace. For Tracks 2 853
and 3, we additionally include qualitative comparisons 854
that demonstrate the functional contribution of the spa- 855
tial and temporal tools, respectively. The temporal tool 856
comparison figure is provided in the main paper (Fig. 3); 857
here we focus on the spatial tool comparison and detailed 858
agent traces. 859

A.6.1. Track 1 Example (Easy, s01) 860
Setup. 861

Scenario s01 (factory), target ID 134 (difficulty: 862
Easy). 863
Initial clue: lower_color = Grey → 11 candi- 864
dates. 865

Table 8. Quality control summary across construction stages.

Stage	Method	Key Metric	Result
Gallery (A.1)	Automated (pose gates)	Rejection modes	<code>fail_no_person</code> dominant
Attributes (A.2)	6-model ensemble + expert review	Inter-model agree.	97.8%
STTG (A.3)	HITL curation + priority labeling	Trust rate	94% (18,151/19,321)
Task GT (A.4)	Algorithmic guarantee	Uniqueness/validity	0 failures
Dialogue (A.4)	Auto metrics + spot-check	ID leakage	0 violations

866 Dialogue.

867 **Witness:** “I remember they were wearing grey
868 pants.”
869 **Agent:** “Can you tell me what color their top was?”
870 **Witness:** “They were wearing a black one.”
871 **Agent:** “How about their hairstyle?”
872 **Witness:** “Their hair was pretty short, about ear-
873 length.”

874 Candidate narrowing.

875 Clue (`lower_color=Grey`): 90 → 11.
876 Turn 1 (`upper_color=Black`): 11 → 3.
877 Turn 2 (`hair_style=Short`): 3 → **1** (ID 134,
878 correct).
879

880 A.6.2. Track 2 Example (Medium, s01)**881 Setup.**

882 Scenario s01 (factory), target ID 3, primary
883 camera c01 (inner warehouse).
884 Zone: `F_WAREHOUSE` (cameras c01, c02,
885 c04, c05).
886 Initial clue: `upper_color = Blue` → 15 candi-
887 dates.

888 Agent trace.

889 Turn 1 (attribute): Agent queries `shoes_type`.
890 Witness responds “dress shoes.” Candidates:
891 15 → 2 (IDs 3, 72).
892 Turn 2 (spatial): Agent asks “Was it deep inside
893 the warehouse, or near the entrance?”
894 Witness responds “Deep inside, near those tall
895 shelves in back.”
896 Person 3 is at c01 (inner warehouse); Person
897 72 is at c02 (parking area).
898 Result: 2 → **1** (ID 3, correct).

899 This example demonstrates the Track 2 pattern: at-
900 tribute questions narrow the pool, and a spatial question
901 resolves the final ambiguity between candidates in differ-
902 ent sub-areas of the same zone.

Spatial tool contribution. Figure 8 compares the full
ARGOS agent with its ablated counterpart (spatial tool
removed) on task T2_s10_92 (Factory, Medium). The full
agent issues a single spatial query that provides an or-
thogonal disambiguation axis beyond appearance. With-
out the spatial tool, the agent must rely exclusively on
attribute questions; however, 9 of 11 attribute queries re-
turn “I’m not sure,” leaving the agent unable to converge.
This illustrates why spatial tool removal causes a 33.8 pp
accuracy drop reported in the main paper (Table 3).

A.6.3. Track 3 Example (Medium, s01)**Setup.**

Scenario s01 (factory), target ID 3.
Two sightings: c05 (warehouse) → c08
(lobby), transition $\approx 8.9s$.
STTG edge c05→c08: TRAVEL, range [7.6s,
20.7s], median 11.2s, $n = 189$.
Initial clue: `upper_color = Blue` → 15 candi-
dates.
Observation: “I saw them in the warehouse a
few minutes in, then almost right after near the
lobby entrance.”

Turn 1: Temporal elimination. The agent calls
`check_temporal` for all 15 candidates:

Candidate(s)	Result	Reason
3, 20, 22, 45, 86	FEASIBLE	Within [7.6, 20.7]s
10, 19, 40, 65, 73	IMPOSSIBLE	TIME_REVERSAL ($\delta < -5s$)
102	IMPOSSIBLE	TOO_SLOW (89.4s \gg 20.7s)
1, 63, 72, 89	ELIMINATED	NOT_PRESENT at one or both cameras

Result: 15 → 5 candidates (10 eliminated: 5
TIME_REVERSAL, 1 TOO_SLOW, 4 NOT_PRESENT).

Turn 2: Attribute. Agent queries `lower_color`. Wit-
ness: “Pretty sure their bottoms were black.” Filter: 5 →
1 (ID 3, correct). TWS = 1.0 (2 turns = τ^*).

This example illustrates the Track 3 pattern: temporal
reasoning eliminates the majority of candidates (10 of
15), and a single attribute question resolves the remaining
ambiguity. A visual comparison of the full agent versus
the ablated variant (temporal tool removed) is provided
in the main paper (Fig. 3), where a single temporal check
eliminates 16 of 19 candidates on a harder task instance
(T3_s01_74).

0 Initial Clue

✓ ARGOS Agent (w/ Spatial)

✗ w/o Spatial Tool

S-Clue: the lobby and entrance area (first floor, inside)

Known attributes: lower_type=Sweatpants

Initial candidates: 10 candidates - [16, 25, ..., 98, 103]

1 Ask Attribute upper_color [10→2]

Q: "What color was their top?"

A: "Blue, I'm pretty sure it was blue."

2 Ask Spatial spatial [2→2]

Q: "Was it in the lobby itself, or more in the connecting area?"

A: "Blue, I'm pretty sure it was blue."

3 Ask Attribute Hair_color, hat [2→2]

A: "I'm not sure about that, I didn't get a good look."

... ..

5 Ask Attribute lower_color [2→1]

Q: "What color were their bottoms?"

A: "Pretty sure it was Pink."

Predict

Target is ID #92 (5 turn → ✓ Correct)

1 Ask Attribute age_style [10→10]

Q: "Could you tell if they were male or female?"

A: "Not sure"

2 Ask Attribute Bag, body_shape, ... [10→10]

A: "I'm not sure about that, I didn't get a good look."

... ..

6 Ask Attribute gender [10→6]

Q: "Could you tell if they were male or female?"

A: "Hmm, Female maybe?"

9 Ask Attribute age_style [6→6]

Q: "Did they look young, middle-aged, or older?"

A: "Not sure"

(11 turn → ✗ Failed)

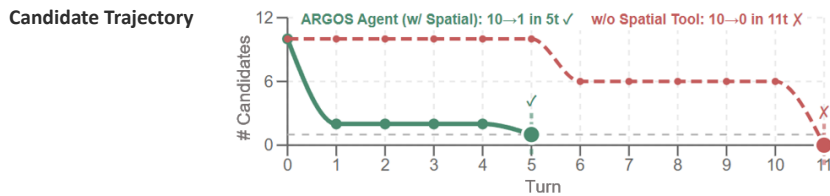


Figure 8. Track 2 tool demonstration: agent with spatial tool (left) vs. without spatial tool (right) on task T2_s10_92 (Factory, Medium). The spatial tool enables disambiguation via location queries, providing an orthogonal axis when appearance attributes are ambiguous.

940 **B. Benchmark Overview Details**

941 This section provides benchmark overview materials
942 omitted from the main paper due to space constraints.

943 **B.1. Three Tracks Overview**

944 Figure 9 illustrates the three progressive tracks of the
945 ARGOS benchmark. All tracks share the same gallery,
946 attribute schema, and scenarios. Track 1 can be evalu-
947 ated independently; Tracks 2 and 3 presuppose Track 1’s
948 semantic parsing capability.

949 **B.2. Benchmark Statistics**

950 Table 9 summarizes the benchmark statistics for each
951 track, including task counts, interaction modes, core ca-
952 pabilities tested, and primary evaluation metrics.

953 **C. Agent Architecture Details**

954 This section provides detailed agent architecture informa-
955 tion that was condensed in the main paper (Sec. 3).

C.1. System Architecture

Figure 10 shows the separation between the **agent** (LLM-
driven) and the **environment** (deterministic).

Environment. The environment consists of two compo-
nents. The *Witness Simulator*, defined in the main paper
(Sec. 2), returns natural-language answers to the agent’s
questions. The *State Manager* provides three determinis-
tic services: a Gallery Service that exposes the attribute
database (read-only), a Filter Service that performs exact-
match attribute filtering and spatial camera mapping, and
a Session Controller that enforces the 20-turn budget and
termination conditions. All State Manager responses are
deterministic and require no LLM reasoning; the agent
selects actions and the environment executes them, fol-
lowing a standard `env.step(action)` paradigm.

Agent. The agent processes each turn through four
modules in sequence:

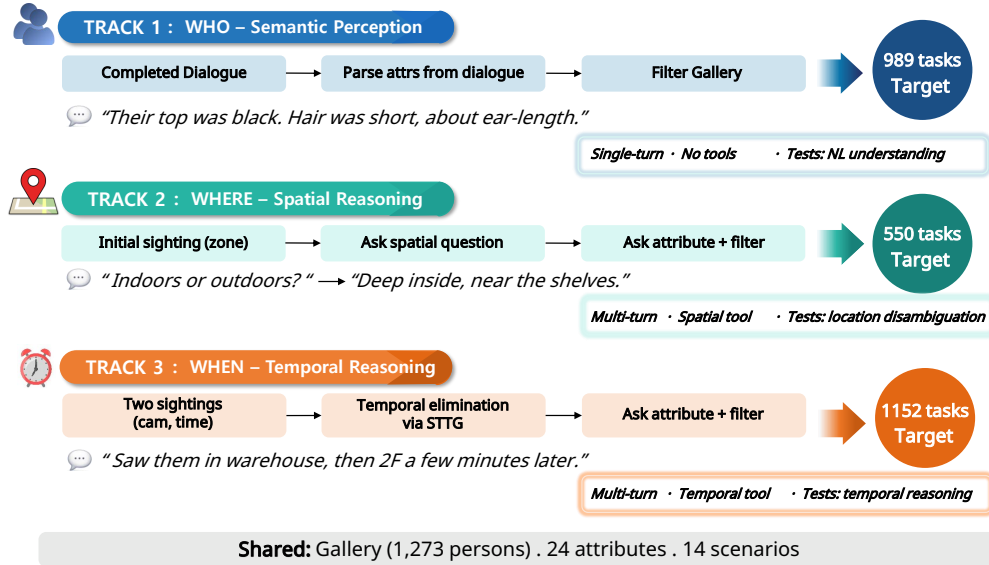


Figure 9. Overview of the three ARGOS tracks. Track 1 receives a completed dialogue for single-turn attribute parsing. Track 2 begins with a zone-level sighting and resolves within-zone ambiguity through spatial and attribute questions. Track 3 provides two temporally separated sightings and eliminates candidates whose transitions violate STTG constraints.

Table 9. ARGOS benchmark statistics across three tracks.

	Track 1 (<i>Who</i>)	Track 2 (<i>Where</i>)	Track 3 (<i>When</i>)
Task count	989	550	1,152
Interaction mode	Single-turn	Multi-turn	Multi-turn
Core capability	Semantic parsing	Spatial reasoning	Temporal reasoning
Avg. dialogue context	5.6 turns	—	—
Avg. oracle turns (τ^*)	—	2.02	1.89
Primary metric	Top-1 Acc	TWS	TWS

- Analyst** queries the gallery, computes attribute distributions over the current candidate set, and retrieves zone structure (Track 2). It identifies which attributes have the highest elimination power among remaining candidates.
 - Planner** receives the Analyst’s summary together with the full dialogue history and decides the next action: ask about an attribute, request a spatial description, or issue a temporal check.
 - Interviewer** executes the chosen action by invoking the appropriate tool. For Track 3, temporal feasibility checking via `check_temporal` (T5) is enforced as the mandatory first action.
 - Interpreter** parses the witness’s natural-language response into a canonical attribute value and applies the corresponding filter to update the candidate set.
- This sequential pipeline runs until the agent issues a prediction or exhausts the turn budget.

C.2. Information Boundary

Figure 11 makes the agent’s information boundary explicit. A critical consequence is that the agent does not know that the witness can only report on three observable attributes (gender, upper-body color, lower-body color) out of 21 in the gallery; if the highest-elimination-power attribute happens to be non-observable, the turn yields no information, forcing the agent to adapt its strategy from failed queries.

C.3. Tool Registry

Figure 12 lists the eight tools available to the agent. Tools T1–T3 support analysis, T4–T5 support interaction with the environment, and T6–T8 produce filtering or prediction actions.

Two aspects of this tool design merit attention. First, T5 (`check_temporal`) provides the agent’s interface to the STTG. Given the current candidate set and a reference timestamp, the State Manager queries the STTG and returns the subset of candidates whose traversal times are feasible. This single call can eliminate a substantial

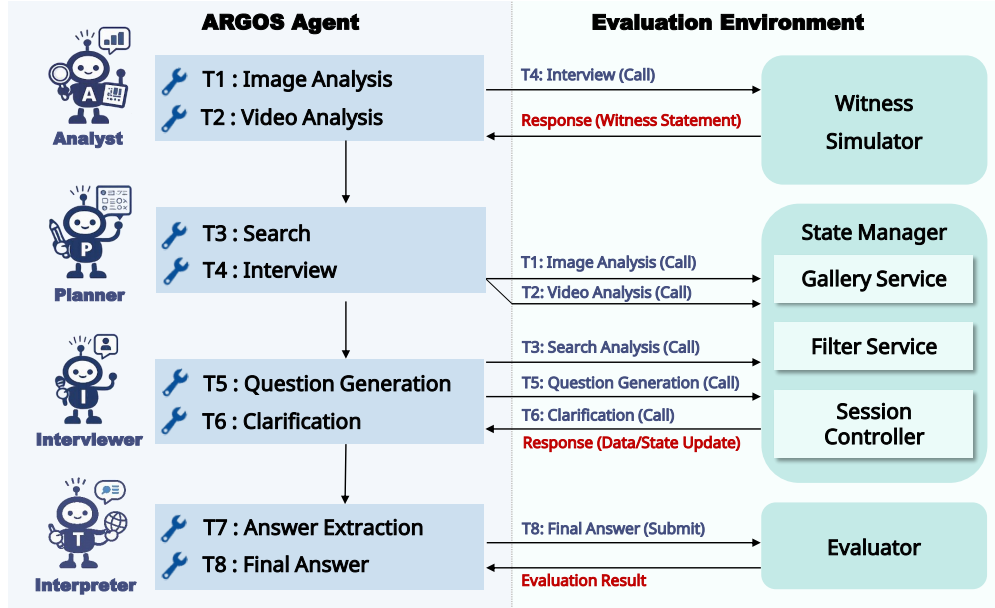


Figure 10. System architecture. **Left:** the ARGOS Agent consists of four LLM-driven modules that form an observe-think-act loop. **Right:** the evaluation environment provides a Witness Simulator and a deterministic State Manager. Arrows denote tool calls (downward) and responses (upward). The agent selects actions; the environment executes them.

✓ Accessible	✗ Not accessible
Gallery DB (21 attributes, read-only)	Candidate-camera mapping
Current candidate ID list	Ground-truth disambiguation path
Zone structure (sub-areas, cameras)	Which attributes the witness can answer
STTG statistics (min, median, max, std)	Per-candidate feasibility details
Available action types	Witness’s internal observable set

Figure 11. Information boundary of the ARGOS Agent. The agent cannot observe which attributes the witness will answer, forcing strategic decisions under uncertainty.

ID	Name	Input	Output	Tracks
T1	query_gallery	candidate IDs	attribute records	1, 2, 3
T2	analyze_dist	candidate IDs	attribute counts, elimination power	1, 2, 3
T3	get_zone_info	—	sub-areas, camera list	2
T4	ask_witness	Question type	NL response (string)	2, 3
T5	check_temp	candidate IDs	filtered partition	3
T6	filter_attr	attr, value, IDs	filtered IDs	1, 2, 3
T7	filter_spatial	camera ID, IDs	filtered IDs	2
T8	predict	candidate ID	correctness flag	1, 2, 3

Figure 12. Tool registry. Each tool is executed by the environment (State Manager or Witness Simulator); the agent invokes them through structured function calls. Track availability indicates which tracks permit each tool. Natural-language parsing at T6 is the primary performance bottleneck.

trivial step where errors directly degrade the candidate set. 1017
1018

D. Additional Experimental Results 1019

This section provides detailed experimental analyses that support and extend the main paper’s findings. 1020
1021

D.1. NL Parsing Analysis 1022

To isolate NL parsing performance from tool-use strategy, we analyze attribute-ask outcomes from the NoSpatial (Track 2) and NoTemporal (Track 3) ablation agents, which rely exclusively on attribute questions and thus maximize the number of parsing events. 1023
1024
1025
1026
1027

Overall parsing statistics. Table 10 summarizes the parsing pipeline. Across both tracks, approximately 85% of witness responses are “I’m not sure” (uncertain), consistent with only 3 of 21 task-generation attributes being visually observable (gender, upper_color, lower_color), yielding a 14.3% expected answer rate. Among answered responses, Track 3 achieves 94.2% parse-and-filter success, while Track 2 reaches only 78.6%. 1028
1029
1030
1031
1032
1033
1034
1035

Per-template breakdown. Table 11 reports parsing success rates for each of the 12 witness response templates plus the catch-all OTHER category. For Track 3, all templates achieve 91–98% success. For Track 2, the 12 structured templates perform comparably (81–90%), but the OTHER category drops to 59.6%. The OTHER 1036
1037
1038
1039
1040
1041

1011 fraction of candidates without any dialogue, making it
1012 the most information-dense tool in the registry. Second,
1013 T6 (*filter_by_attribute*) takes as input a canonical
1014 attribute value that the agent must extract from the
1015 witness’s natural-language response. Because the witness
1016 uses varied phrasings, natural-language parsing is a non-

Table 10. NL parsing statistics from ablation agents (attribute-only mode).

Metric	Track 2 (NoSpatial)	Track 3 (NoTemporal)
Total attribute asks	8,587	18,006
Witness answered	1,309 (15.2%)	2,633 (14.6%)
Witness uncertain	7,278 (84.8%)	15,373 (85.4%)
Parse+filter success	1,029 (78.6% of ans.)	2,481 (94.2% of ans.)
No reduction after parse	280 (21.4% of ans.)	152 (5.8% of ans.)

Table 11. Per-template parsing success rate. The OTHER category (GPT-generated free-form responses) drives Track 2’s lower overall rate.

Template	Track 3			Track 2		
	Total	Succ.	Rate	Total	Succ.	Rate
OTHER (GPT-generated)	448	417	93.1%	361	215	59.6%
“Pretty sure it was X.”	233	218	93.6%	89	72	80.9%
“Not totally sure, but X-ish.”	201	183	91.0%	84	73	86.9%
“It looked like X.”	188	177	94.1%	77	64	83.1%
“They had X.”	186	172	92.5%	88	75	85.2%
“From what I saw, X.”	185	179	96.8%	69	60	87.0%
“I noticed X.”	178	171	96.1%	92	81	88.0%
“I believe it was X.”	178	167	93.8%	82	74	90.2%
“Something like X, I think.”	175	168	96.0%	69	59	85.5%
“Hmm, X maybe?”	173	163	94.2%	78	70	89.7%
“They definitely had...”	168	164	97.6%	76	64	84.2%
“I think it was X.”	160	152	95.0%	80	65	81.2%
“I’d say X.”	160	150	93.8%	64	57	89.1%

Table 13. Top-1 accuracy (%) by difficulty level across tracks and settings. Shaded cells indicate difficulty inversions (higher difficulty, higher accuracy).

Track / Setting	Model	Easy	Medium	Hard
<i>Track 2 — Main Experiment</i>				
	GPT-5.2	89.9	72.3	53.8
	GPT-4o	92.6	73.8	53.8
	GPT-5-mini	92.6	73.8	55.5
	Claude Sonnet 4	92.6	75.9	55.5
<i>Track 3 — Main Experiment</i>				
	GPT-5.2	95.2	88.5	84.3
	GPT-4o	76.6	82.2	79.0
	GPT-5-mini	94.5	88.5	83.9
	Claude Sonnet 4	82.1	83.8	83.9
<i>Track 2 — InitialOnly Baseline</i>				
	GPT-4o	62.4	64.2	68.9
	GPT-5-mini	64.4	69.1	83.2

processes all dialogue turns at once and extracts attribute-value pairs from free-form text. Each parsing error can eliminate the correct candidate from the filtered set. With multiple turns of dialogue, errors accumulate, making the complete dialogue less reliable than the initial statement alone.

In contrast, Track 3 shows the expected pattern (FullDialogue > InitialOnly for both models), because Track 3’s parsing success rate is 94.2% and the temporal structure provides strong disambiguation even with occasional parsing errors.

This inversion highlights the need for tool-augmented interactive parsing, where the agent processes one response at a time with structured tool calls, rather than batch-processing entire dialogues. The interactive agent (LLM-Agent) achieves 74.5% (GPT-4o) on Track 2, substantially above both FullDialogue and InitialOnly.

For completeness, Track 3 results: GPT-4o FullDialogue 29.6% vs. InitialOnly 11.3%; GPT-5-mini FullDialogue 29.9% vs. InitialOnly 11.5%. Both show the expected direction, confirming that the inversion is Track 2-specific and attributable to the higher parsing difficulty of spatial dialogue.

D.3. Difficulty-Level Analysis

Table 13 reports per-difficulty performance for all models and key ablation settings. Two observations stand out.

Track 3 Easy inversion (GPT-4o, Claude Sonnet 4). GPT-4o achieves only 76.6% on Easy tasks, below its Medium (82.2%) and Hard (79.0%) scores. A similar pattern appears with Claude Sonnet 4 (Easy 82.1% vs. Medium 83.8%). Track 3 Easy tasks have few attribute candidates (≤ 6) and at least two temporal eliminations. While the temporal reasoning overhead is minimal for

template captures conversational responses generated by GPT during dialogue wrapping (Stage 2), which introduce varied phrasing, hedging, and contextual elaboration that the rule-based parser struggles to normalize.

This gap is the primary driver of the 78.6% vs. 94.2% disparity between tracks. Track 2 dialogues, generated to include spatial context and location descriptions, produce more diverse natural language patterns (OTHER accounts for 27.6% of Track 2 answered responses vs. 17.0% for Track 3).

D.2. FullDialogue vs InitialOnly Inversion (Track 2)

In Track 2, the FullDialogue baseline (which feeds the entire ground-truth dialogue to the LLM in one pass, bypassing interactive tools) performs *worse* than InitialOnly (which provides only the initial witness statement). This counterintuitive inversion is reproduced across two models:

Table 12. FullDialogue vs. InitialOnly (Track 2). More dialogue context hurts performance.

Model	FullDialogue	InitialOnly	Δ
GPT-4o	56.2%	64.7%	-8.5pp
GPT-5-mini	66.4%	70.9%	-4.5pp

The explanation connects directly to the parsing analysis in Sec. D.1: Track 2’s 78.6% parsing success rate means that roughly one in five answered attributes is parsed incorrectly. In the FullDialogue setting, the LLM

Table 14. Observable attribute (K_3) resolution rate per scenario. Only 35.3% of tasks are resolvable by appearance attributes alone.

Scen.	Resolved	Total	Rate	Scen.	Resolved	Total	Rate
s01	30	85	35.3%	s34	29	87	33.3%
s10	30	87	34.5%	s35	30	78	38.5%
s11	28	81	34.6%	s36	27	77	35.1%
s13	23	87	26.4%	s38	36	77	46.8%
s16	23	84	27.4%	s39	31	90	34.4%
s18	26	66	39.4%	s42	21	93	22.6%
s20	31	79	39.2%	s47	42	81	51.9%

Weighted mean: 35.3% Range: 22.6% (s42) – 51.9% (s47)

these tasks, the small candidate pool leaves little room for parsing error: a single misclassified attribute can eliminate the target. GPT-5.2 and GPT-5-mini, with higher overall parsing accuracy, do not exhibit this inversion.

Track 2 InitialOnly complete inversion. Both GPT-4o and GPT-5-mini show a monotonic reversal where Hard tasks score highest (68.9% and 83.2%, respectively). This pattern is absent from the interactive LLM-Agent (which shows the expected Easy > Medium > Hard ordering). The explanation lies in the difficulty definition: Hard tasks have larger initial candidate pools and longer disambiguation paths, requiring more interactive turns. However, the InitialOnly setting provides only the initial witness statement with no follow-up. Hard tasks receive a richer initial description (covering more attributes in the opening statement), which paradoxically provides more information for a single-pass prediction. This inversion reinforces that the difficulty labels measure *interactive* complexity, and the benchmark genuinely requires multi-turn dialogue to follow the intended difficulty gradient.

D.4. Observable Attribute Resolution Rate

The information boundary (main paper Sec. 4.1) defines which attributes are visually observable by the witness simulator. Using the three observable attributes $K_3 = \{\text{gender}, \text{upper_color}, \text{lower_color}\}$, we measure the resolution rate: the fraction of tasks where K_3 alone can uniquely identify the target.

Table 14 reports per-scenario resolution rates. The weighted mean is 35.3% (range: 22.6%–51.9%), meaning that approximately 65% of tasks cannot be solved by observable attributes alone and require the agent to use spatial or temporal tools. This validates the benchmark’s design: the information boundary ensures that multi-modal reasoning (beyond appearance) is necessary for the majority of tasks.

D.5. Extra Experiments: GPT-5.2 High-Setting

To test whether increased compute budget improves performance, we evaluate GPT-5.2 with enhanced settings

Table 15. GPT-5.2 (high reasoning) vs. GPT-4o (default) on two scenarios.

Track	Scenario	GPT-5.2 (high512)			GPT-4o (base)		
		n	Top-1	TWS	n	Top-1	TWS
T2	s01 (Factory)	43	76.7%	0.356	43	88.4%	0.396
T2	s39 (University)	48	72.9%	0.299	48	87.5%	0.336
T3	s01 (Factory)	85	84.7%	0.572	85	87.1%	0.541
T3	s39 (University)	90	91.1%	0.597	90	74.4%	0.526

(reasoning effort: high, max output tokens: 512) on two representative scenarios (s01, factory; s39, university) and compare against GPT-4o with default settings.

The results in Table 15 show no consistent advantage for the high-compute setting. GPT-5.2 (high512) outperforms GPT-4o on Track 3 s39 (91.1% vs. 74.4%) but underperforms on Track 2 for both scenarios. The mixed results suggest that the benchmark’s difficulty stems from structural challenges (NL parsing, spatial reasoning, temporal inference) rather than reasoning depth, and that simply scaling inference-time compute does not overcome these bottlenecks.

D.6. Metric Details

Turn-Weighted Success (TWS). For tasks where the agent correctly identifies the target (correct-only), TWS rewards faster resolution:

$$\text{TWS}_i = \frac{\tau_i^*}{\max(\tau_i^*, t_i)}, \quad (5)$$

where τ_i^* is the oracle optimal turn count and t_i is the agent’s actual turn count. $\text{TWS} = 1.0$ when the agent matches the oracle; $\text{TWS} < 1.0$ when the agent takes extra turns. For incorrect predictions, $\text{TWS} = 0$. The benchmark-level TWS is the average across all tasks.

AUC-CRR (Area Under Candidate Reduction Rate).

Unlike TWS, AUC-CRR evaluates all tasks regardless of correctness. It measures the cumulative candidate reduction efficiency over turns:

$$\text{AUC-CRR}_i = \frac{1}{\tau_i^*} \sum_{t=1}^{\tau_i^*} \left(1 - \frac{|C_t|}{|C_0|}\right). \quad (6)$$

Because AUC-CRR includes incorrect tasks, it can produce different model rankings than TWS. For example, in Track 3, GPT-4o ranks 2nd by TWS (0.567) but 4th by AUC-CRR (0.920), because its higher premature prediction rate inflates the turn count for correct tasks while the candidate reduction on incorrect tasks is penalized differently.

Oracle turn count (τ^*) computation. τ^* is determined by the ground-truth disambiguation path length

Table 16. Agent behavioral statistics (Track 2 / Track 3, all four models).

Metric	Track 2				Track 3			
	5.2	4o	5m	CS4	5.2	4o	5m	CS4
Over-filter (%)	0.3	0.3	0.3	0.3	0.5	0.4	0.3	0.3
Redundant Q (%)	0	0	0.7	0	0	0	0.2	0
Wrong Tool	0	0	0	0	0	0	0	0
Timeout (%)	3.1	3.3	4.2	3.3	5.4	1.6	3.8	1.4
Premature Pred.	70	70	73	68	64	212	90	163

Table 17. Cross-model failure distribution (4 models evaluated per task).

Category	Track 2	Track 3
All correct (0/4 failed)	381 (69.3%)	794 (68.9%)
1/4 failed	19 (3.5%)	164 (14.2%)
2/4 failed	13 (2.4%)	99 (8.6%)
3/4 failed	35 (6.4%)	56 (4.9%)
Universal failure (4/4 failed)	102 (18.5%)	39 (3.4%)

Table 18. Universal failure rate by difficulty level.

Difficulty	Track 2			Track 3		
	Failed	Total	Rate	Failed	Total	Rate
Easy	7	149	4.7%	1	145	0.7%
Medium	51	282	18.1%	24	702	3.4%
Hard	44	119	37.0%	14	305	4.6%

Table 19. Failure reason patterns in universal failure cases.

Track	Pattern	Count	% of universal
T2	All 4 wrong prediction	86	84.3%
T2	All 4 timeout	13	12.7%
T2	Mixed	3	2.9%
T3	Mixed (1 timeout + 3 wrong)	17	43.6%
T3	All 4 timeout	15	38.5%
T3	All 4 wrong prediction	4	10.3%
T3	Other mixed	3	7.7%

tural: when a task is hard for one model, it tends to be hard for all. 1210
1211

Difficulty correlation. Universal failure rates scale with difficulty in both tracks (Figure 13(b); Table 18). Track 2 Hard tasks have a 37.0% universal failure rate, confirming that the difficulty labels capture genuine challenge levels. 1212
1213
1214
1215
1216

Failure patterns. Table 19 breaks down the failure reasons for universal cases. In Track 2, 84.3% of universal failures are wrong predictions (all four models predict an incorrect person) and only 12.7% are timeouts. In Track 3, the pattern reverses: 38.5% are all-timeout and 43.6% are mixed (some timeout, some wrong prediction). 1217
1218
1219
1220
1221
1222

Candidate reduction analysis. As illustrated in Figure 13(d), among Track 2 universal failures (GPT-4o), 50% (51 out of 102) narrowed the candidate pool to exactly one person but chose the wrong one. This confirms that NL response parsing, not search strategy, is the primary bottleneck: the agent successfully navigates the disambiguation path but makes a parsing error at a critical turn, eliminating the correct candidate and converging on an incorrect one. In contrast, Track 3 universal failures show higher residual candidate counts (mean 3.0–3.2 1223
1224
1225
1226
1227
1228
1229
1230
1231
1232

constructed during task generation (Sec. A.4). For Track 1, τ^* equals the number of IG simulation turns. For Track 2, $\tau^* = \text{spatial_turns} + \text{attribute_turns}$. For Track 3, $\tau^* = 1 + \text{attribute_turns}$ (one temporal check plus attribute disambiguation).

Premature prediction. A premature prediction occurs when the agent issues a `predict` action while $|C_{\text{final}}| > 1$ and the prediction is incorrect. This measures overconfident behavior where the agent guesses before resolving ambiguity. Premature prediction rates range from 11.8–18.3% across models on Track 2 (68–73 occurrences out of 550 tasks) and 5.6–18.4% on Track 3.

D.7. Agent Behavioral Statistics

Table 16 reports behavioral statistics across all four models.

Three categories of tool misuse (over-filtering, redundant questions, wrong tool usage) are near zero across all models, confirming that the structured tool interface effectively constrains agent behavior. The primary failure modes are timeout (reaching the maximum turn limit) and premature prediction. GPT-4o shows an elevated premature prediction count on Track 3 (212 out of 1,152 tasks), suggesting that it tends to predict before fully disambiguating, which contributes to its lower TWS despite reasonable Top-1 accuracy.

D.8. Failure Case Analysis

We analyze cases where all four evaluated models (GPT-5.2, GPT-4o, GPT-5-mini, Claude Sonnet 4) fail on the same task, which we term *universal failures*. These represent structurally difficult tasks that current LLMs cannot solve regardless of model choice. Figure 13 provides a visual overview of the failure landscape; Tables 17–19 report the exact statistics.

Prevalence. As shown in Figure 13(c) and Table 17, both tracks share a similar all-correct rate (~69%), but Track 2 concentrates its failures in the universal category (18.5% vs. 3.4% for Track 3), while Track 3 distributes failures more evenly across partial categories. This indicates that Track 2’s spatial reasoning failures are *struc-*

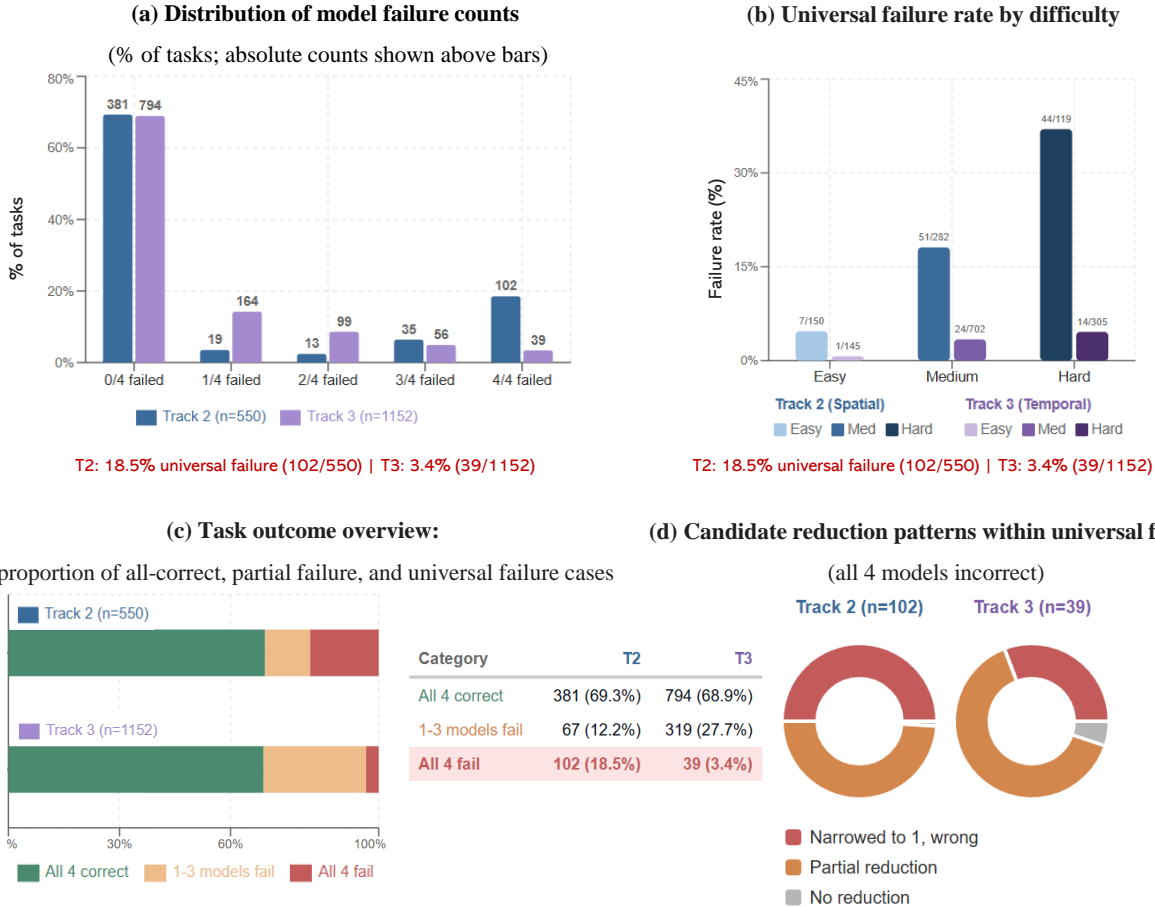


Figure 13. **Universal failure analysis.** (a) Distribution of per-task model failure counts. (b) Universal failure rate by difficulty level. (c) Task outcome overview (all-correct, partial, universal). (d) Candidate reduction patterns in Track 2 universal failures: 50% narrow to exactly 1 candidate but predict incorrectly.

1233 across models), indicating that the temporal reasoning
1234 itself is incomplete rather than a final-stage parsing error.

1235 **D.9. Qualitative Analysis: Strategy and Parsing**
1236 **Gap**

1237 This section presents qualitative case studies on question-
1238 selection strategy and the NL parsing gap between oracle
1239 and agent. Qualitative demonstrations of the spatial and
1240 temporal tools are in Sec. A.6 (Figure 8) and the main
1241 paper (Fig. 3).

1242 **Question-selection strategy (Figure 14).** On task
1243 T2.s20.32 (Factory, Medium), the LLM-guided
1244 information-gain strategy selects `lower_color` as its
1245 first question—the most discriminative attribute, reduc-
1246 ing candidates from 9 to 3 in a single turn—and resolves
1247 the task in 3 turns total. The random-ordering baseline
1248 begins with `shoes_color`, which yields no candidate
1249 reduction, and wastes 7 turns on uninformative attributes
1250 before eventually reaching the same discriminative ques-

tion. Both agents ask the same set of questions; the
difference is entirely in ordering, highlighting that strate-
gic attribute selection is as important as the attributes
themselves.

1251 **Oracle vs. agent parsing gap (Figure 15).** Finally,
1252 we visualize the NL parsing bottleneck identified in
1253 Sec. D.8. On task T2.s42.21 (School, Hard), both the
1254 oracle and the LLM agent ask the same first question
(`upper_type`). The oracle receives a structured value
1255 “Jacket” and performs exact-match filtering, resolving
1256 the task in 2 turns. The agent receives the natural-
1257 language response “a black jacket”; a slight parsing dif-
1258 ference causes a mismatch in the attribute filter, leading
1259 to 7 stalled turns with no candidate reduction and ulti-
1260 mate convergence to the wrong candidate. This single-
1261 question divergence—identical input, different parsing—
1262 encapsulates why NL parsing accuracy is the primary
1263 bottleneck for future improvement and motivates the de-
1264 velopment of robust attribute extraction modules.
1265
1266
1267
1268
1269

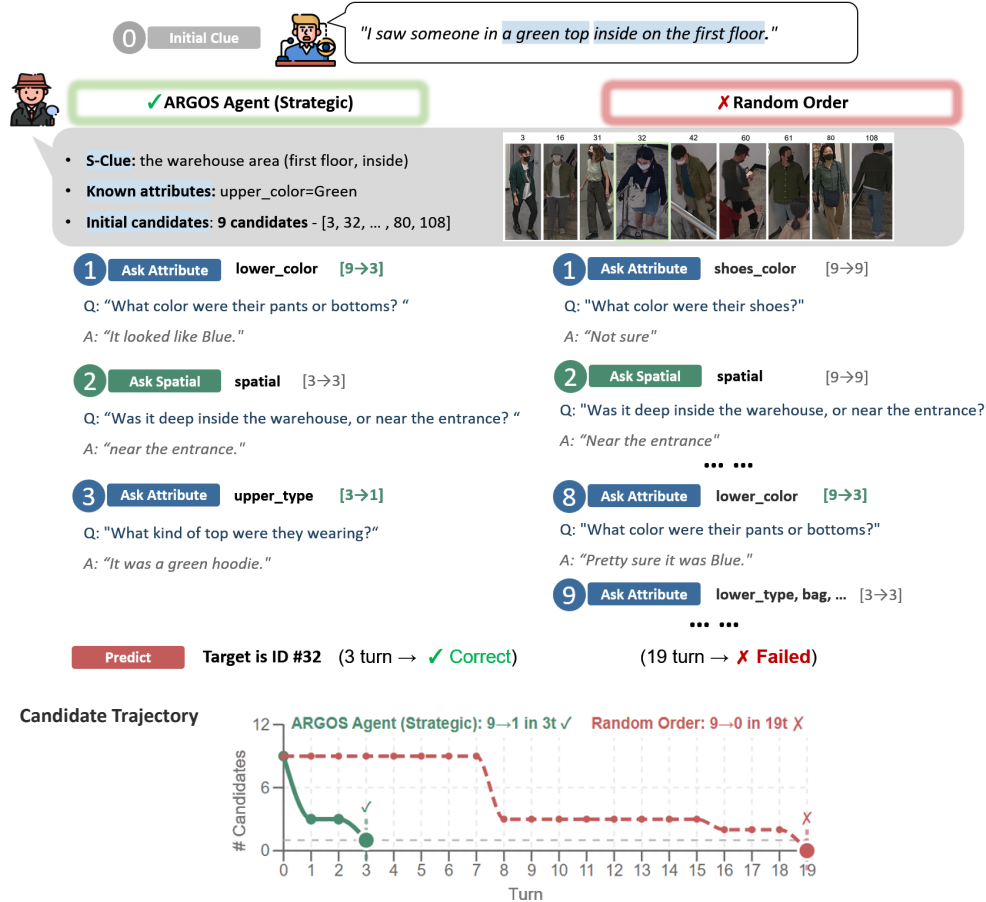


Figure 14. **Qualitative comparison: ARGOS agent with LLM-guided strategy (left) vs. random question ordering (right) on task T2_s20_32 (Factory, Medium).** Strategic attribute selection resolves the task in 3 turns; random ordering wastes 7 turns on uninformative attributes before reaching the same discriminative question.

1270 **E. Additional Benchmark Details**

1271 **E.1. Full Attribute Taxonomy**

1272 Table 20 lists all 24 attributes in the ARGOS gallery
 1273 database. Attributes are organized into five categories:
 1274 head (hair and accessories), upper body (garment, bag),
 1275 lower body (garment), feet and hands, and soft bio-
 1276 metrics. Each attribute has a predefined set of possi-
 1277 ble values; multi-select attributes (upper_state,
 1278 body_features) can hold multiple values per person.
 1279 Three attributes (hair_visibility, leg_visibility,
 1280 body_features) are excluded from the IG simulation during task generation
 1281 due to extreme class imbalance, but remain in the
 1282 gallery database accessible to agents. Two additional
 1283 attributes (shoe_color, visual_age_style) are
 1284 excluded from initial clue selection but remain available
 1285 for IG-based questioning during disambiguation. The
 1286 gallery identifier (global_id) is a record key, not a
 1287 visual attribute, and is not used in the IG simulation.
 1288

E.2. Witness Simulator Templates

The witness simulator generates natural-language re-
 sponses to agent queries using 12 response templates
 and a deterministic observable-attribute rule. Table 21
 lists all templates.

Response logic. The witness responds according to
 three priority rules: (i) If the queried attribute is on the
 ground-truth disambiguation path (constructed during
 task generation, Sec. A.4), the witness always answers
 using the pre-generated conversational response from
 Stage 2 dialogue wrapping. (ii) If the attribute is out-
 side the disambiguation path but within the observable
 set $K_3 = \{\text{gender, upper_color, lower_color}\}$,
 the witness answers with the ground-truth value wrapped
 in a randomly selected template (seed = 42 for repro-
 ducibility). (iii) For all other attributes, the witness re-
 sponds "I'm not sure about that, I didn't get a good look."

This design ensures that agents cannot bypass the
 intended disambiguation challenge by asking about non-



Figure 15. **Oracle (ground-truth attributes) vs. ARGOS agent on task T2.s42.21 (School, Hard)**. The oracle resolves in 2 turns with exact-match filtering. The agent parses NL responses with slight errors, causing 7 stalled turns and convergence to the wrong candidate. This gap highlights NL parsing accuracy as the primary bottleneck for future improvement.

1308 observable attributes, while still receiving informative
1309 responses for attributes that a real witness could plausibly
1310 recall. The 85% uncertain rate observed in experiments
1311 (Sec. D.1) is consistent with 3 of 21 task-generation at-
1312 tributes being observable.

1313 **Example.** If the agent asks about `upper_color` (ob-
1314 servable, $\in K_3$) and the target wears a blue top, the
1315 witness might respond: "Pretty sure it was Blue." If the
1316 agent asks about `shoes_type` (not in K_3 and not on
1317 the disambiguation path), the witness responds: "I'm not
1318 sure about that, I didn't get a good look."

1319 **Spatial and temporal actions.** In Track 2, the agent
1320 can also perform an `ask_spatial` action (at most once
1321 per task), to which the witness provides a pre-generated
1322 location description from the zone disambiguation tree
1323 (e.g., "Deep inside, near those tall shelves in back").
1324 In Track 3, the `check_temporal` action triggers the
1325 temporal feasibility check using pre-computed STTG-

based classifications; this action does not involve the
1326 witness but returns structured feasibility results for all
1327 candidates.
1328

E.3. Environment STTG Details 1329

This section provides the full STTG visualizations and
1330 zone definitions for both environments. The factory
1331 STTG is also shown in the main paper (Fig. 2); here
1332 we present both environments side by side for compari-
1333 son.
1334

Factory environment. The factory STTG contains 110
1335 directed edges across 9 atomic zones and 2 composite
1336 zones (Table 22). The environment spans two floors
1337 connected by multiple stairwells, with the warehouse
1338 area (4 cameras) forming the largest atomic zone.
1339

University campus. The university STTG contains 149
1340 directed edges across 6 atomic zones and 2 composite
1341 zones (Table 23). The campus features a central building
1342

Table 20. Complete attribute taxonomy (24 attributes). †: excluded from IG simulation. ‡: excluded from initial clue selection. All include “Uncertain” as a possible value.

Category	Attribute	Possible Values
Head: Hair	hair_visibility†	Visible, Covered by Hat, Bald/Shaved
	hair_style	Short (Ear-length), Medium (Neck-length), Long (Shoulder+), Ponytail/Bun, Afro/Textured, Bald, Other
	hair_color	Black, Brown, Blonde, Gray/White, Dyed/Unnatural, Other
Head: Acc.	headwear_type	None, Cap, Beanie, Helmet, Bucket Hat, Other
	eyewear_type	None, Glasses, Sunglasses, Other
	mask_state	Properly Worn, Chin Mask, No Mask, Other
Upper	upper_garment_type	T-shirt, Shirt, Hoodie, Sweatshirt, Jacket, Vest, Suit, Puffer/Padding (Long/Short), Coat, Dress, Uniform/Gown, Other
	upper_color_layout	Solid, Layered, Patterned, Colorblock, Graphic/Logo, Other
	upper_garment_color	Black, White, Grey, Red, Blue, Green, Yellow, Orange, Purple, Pink, Brown, Neon, Other
	upper_state (multi)	Hooded, Zipper Open/Closed, Long/Short Sleeve, Sleeveless, Collared, None
	upper_fit_style	Regular, Loose/Oversized, Tight/Fitted, Bulky (Padding)
	torso_bag_type	None, Backpack, Shoulder Bag, Crossbody Bag, Lanyard/ID, Other
Lower	leg_visibility†	Fully Visible, Partially Covered, Hidden
	lower_garment_type	Trousers, Jeans, Sweatpants, Shorts, Skirt, Leggings, Work/Cargo Pants, Other
	lower_garment_color	Black, White, Grey, Red, Blue, Green, Yellow, Orange, Purple, Pink, Brown, Neon, Other
	lower_fit_style	Regular, Baggy/Loose, Skinny/Tight, Short
Feet/Hands	shoe_type	Sneakers, Boots/Walker, Dress Shoes, Sandal/Slipper, Other
	shoe_color‡	(same 14 colors as upper/lower)
	items_held	None, Phone, Bag/Carrier, Box, Notebook, Umbrella, Drink, Tool, Paper, Other
Soft Bio	visual_age_style‡	Child/Teen, Young Adult, Mature, Elderly, Uniformed
	body_shape	Slender, Normal, Heavy
	body_features† (multi)	Potbelly, Stocky, Muscular, Thick Thighs, Obese, Stick-like Limbs, Lanky, Petite, Thin Wrists, Frail, Tall, Short, Other, None
	visual_gender	Male, Female

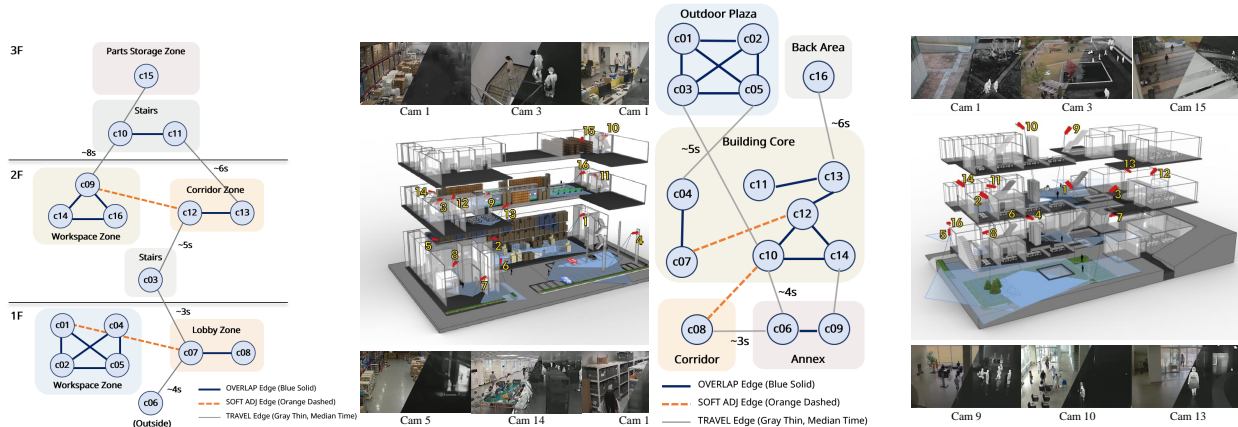


Figure 16. **STTG comparison.** *Left:* Factory environment (16 cameras, 110 edges, 9 atomic zones). *Right:* University campus (16 cameras, 149 edges, 6 atomic zones). Edge types: OVERLAP (blue), SOFT_ADJ (orange), TRAVEL (gray). The university environment has denser inter-zone connectivity due to its open outdoor layout.

1343 core (6 cameras) surrounded by outdoor plazas and paths,
 1344 resulting in denser inter-zone connectivity compared to
 1345 the factory’s vertically separated floor plan.

1346 **F. Problem Scope and Emergent Challenges**

1347 This section clarifies the key characteristics of our prob-
 1348 lem formulation and discusses the new challenges that
 1349 emerge compared with the closest existing problem set-
 1350 tings.

F.1. Key Features of Our Problem Formulation

[Content from ECCV supplementary Sec. D.1: three key features (interactive chat-based search, inclusion of spatio-temporal clues, ambiguity in witness responses) with detailed discussion of each. The illustration figure (new_teaser_v2.pdf) is provided in the main paper as Fig. 1.]

F.2. Emergent Challenges

[Content from ECCV supplementary Sec. D.2: discussion of how the formulation shifts from similarity learning

Table 21. Witness response templates (12 total). Each template wraps the ground-truth attribute value $\{value\}$ in varied natural language. Templates are selected uniformly at random with a fixed seed.

#	Template
1	“I think it was $\{value\}$.”
2	“Pretty sure it was $\{value\}$.”
3	“I’d say $\{value\}$.”
4	“I noticed $\{value\}$.”
5	“They had $\{value\}$.”
6	“It looked like $\{value\}$.”
7	“I believe it was $\{value\}$.”
8	“From what I saw, $\{value\}$.”
9	“Hmm, $\{value\}$ maybe?”
10	“Something like $\{value\}$, I think.”
11	“Not totally sure, but $\{value\}$ -ish.”
12	“They definitely had some kind of $\{value\}$ going on.”

1361 *to reasoning, requiring iterative information-gain opti-*
1362 *mization under witness uncertainty, heterogeneous evi-*
1363 *dence integration across visual/spatial/temporal modali-*
1364 *ties, and robust operation under incomplete feedback.]*

Table 22. Factory zone definitions. Composite zones merge adjacent atomic zones connected by SOFT_ADJ edges.

Zone ID	Cameras	Name	Type
F_WAREHOUSE	c01, c02, c04, c05	Warehouse area	Atomic
F_LOBBY	c07, c08	Lobby and entrance area	Atomic
F_WORKSPACE	c09, c14, c16	2F work floor	Atomic
F_CORRIDOR_2F	c12, c13	2F corridor and office	Atomic
F_STAIRS_MAIN	c03	Main stairwell	Singleton
F_PASSAGE	c06	Outside the entrance	Singleton
F_STAIRS_SEC	c11	Secondary stairwell	Singleton
F_STAIRS_UPPER	c10	Upper stairwell	Singleton
F_PARTS_STORAGE	c15	Parts storage room	Singleton
FC_1F_INDOOR	c01, c02, c04, c05, c07, c08	1F warehouse + lobby	Composite
FC_2F	c09, c12–c14, c16	2F workspace + corridor	Composite

Table 23. University campus zone definitions. Composite zones merge adjacent atomic zones connected by SOFT_ADJ edges.

Zone ID	Cameras	Name	Type
S_PLAZA	c01, c02, c03, c05	Outdoor plaza	Atomic
S_BUILDING_CORE	c04, c07, c11–c14	Inside the building	Atomic
S_OUTDOOR_PATH	c06, c15	Outdoor path area	Atomic
S_ANNEX	c09, c10	2F main lobby	Atomic
S_CORRIDOR_1F_B	c08	Corridor near entrance	Singleton
S_BACK_AREA	c16	Outdoor area at back	Singleton
SC_INDOOR	c04, c07–c14	Building + corridor + annex	Composite
SC_OUTDOOR	c01–c05, c06, c15, c16	Plaza + paths + back	Composite