# LEARNING IN-THE-WILD: SELECTIVE IN-CONTEXT WAM ALIGNMENT WITH INTERNET-SCALE HUMAN VIDEO RETRIEVAL

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Human videos offer rich manipulation knowledge, but useful procedures are scattered across long, unstructured recordings. References must show the right operation, and policies must learn what transfers across embodiments. We present WildWAM, which combines RIVER for robot-guided Internet video retrieval with SAIL for selective human-to-robot in-context alignment. RIVER localizes and verifies operations through complementary search routes and event-aware windows; SAIL aligns human context with robot history through Internet pre-adaptation. Across 84 retrieval configurations and seven comparison methods, RIVER achieves 83.3% coverage and 62.0% functional precision. A three-corpus index of 715,565 human intervals supports over 100k annotated human–robot pairs; Open-AoE supplies 66.6% of distinct retrieved human clips. A unified 59,589-episode G1 foundation is trained for 350,000 updates. Internet pre-adaptation reduces downstream action-prediction error by 17.2% and dynamic-region MAE by 4.20% over target-only adaptation. Across five real-robot tasks with 15 trials per task and policy which include both single-arm and bimanual manipulations, WildWAM achieves 44.0% full-task success conditioned on human context, compared with 28.0% for target-only adaptation. These results connect diverse in-the-wild human procedures to reference-conditioned robot learning, demonstrating the potential of learning from the unstructured human videos.

## 1 INTRODUCTION

A robot closing a laptop must contact its lid, rotate it toward the base, and complete the closure. A task-focused collection may lack it, but an in-the-wild electronics-repair video may demonstrate exactly this operation. Internet video offers broader procedural coverage; using it requires finding the operation within disorganized recordings and transferring what a different embodiment can use.

Teleoperation provides action-aligned supervision but is expensive to repeat across tasks and environments. Human video reveals object roles, state changes, and procedural order but is too massive to locate precisely matching results. Recent policies exploit curated or generated human–robot correspondences (Jain et al., 2024; Chen et al., 2026; Zhou et al., 2026a). In the wild, relevant motion is brief and cluttered; a visually similar clip may reverse an operation or omit its conclusion.

To this end, We propose **WildWAM**, which combines **RIVER** (Robot-guided Internet Video Event Retrieval) and **SAIL** (Selective Alignment for In-context Learning). RIVER localizes and verifies operations with consistent roles, direction, and outcomes. Correspondence labels distinguish full procedures from useful partial matches. SAIL preserves direct human visual context while using robot history to select transferable content. Internet pre-adaptation prepares the WAM to learn from collected human demonstrations, which specify tasks without external text prompts.

Our contributions are: **(i)** event-complete robot-to-human retrieval with procedural verification, improving matching across candidate scopes; **(ii)** applicability-aware world-action learning with direct reference tokens and history-conditioned reading, yielding downstream action-prediction gains; and **(iii)** a unified robot foundation and annotated Internet pairing collection that connect diverse human procedures to multiple robot embodiments. We will release the unified robot data and pairing manifests with temporal provenance, together with their automatic annotations.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

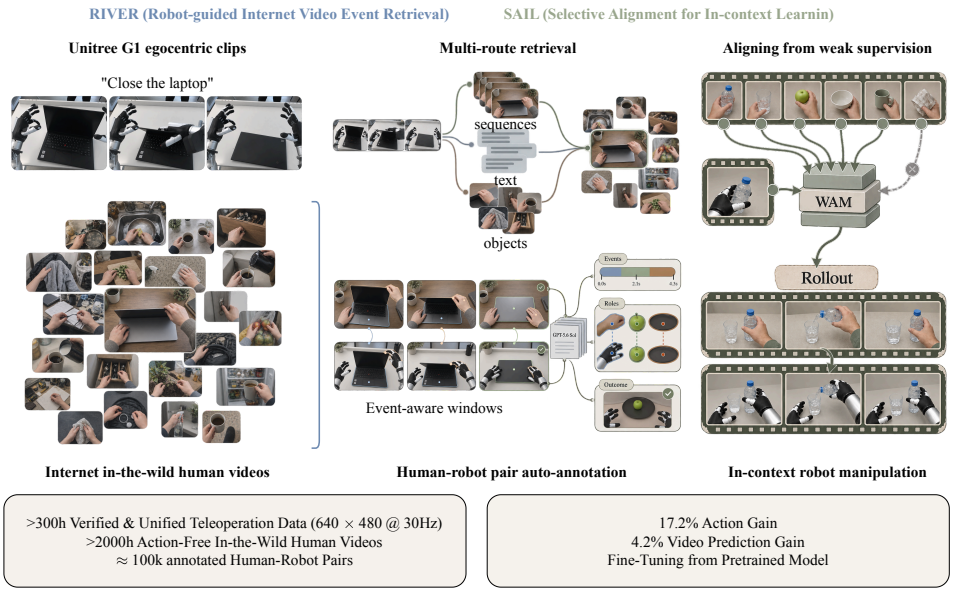



Figure 1: **WildWAM links Internet procedures to robot learning.** RIVER uses robot demonstrations to retrieve human clips through complementary visual and textual routes, select event-complete windows, and verify object roles and outcomes. Pair annotations distinguish full procedures from useful partial correspondences. SAIL uses causal robot history to select relevant human context for joint action and future-video prediction. Internet pre-adaptation is followed by adaptation with collected human demonstrations; at deployment, a human reference specifies the task while model parameters remain fixed.

## 2 Related Work

**Human-video conditioning and robot ICL.** Human video can supervise shared policies or specify tasks at inference. EgoMimic co-trains human and robot data (Kareer et al., 2025); Vid2Robot aligns video with control (Jain et al., 2024). RHyME composes retrieved clips through sequence-level optimal transport to address execution mismatch (Kedia et al., 2025). HOST couples demonstrated progress to robot prediction, Zero-WAM uses generated pairs and future-chunk supervision, and MimicDroid mines human-play behaviors for humanoid ICL (Chen et al., 2026; Zhou et al., 2026a; Shah et al., 2026). WildWAM handles retrieved, partially applicable demonstrations through correspondence supervision and history-conditioned reading.

**World-action models and adaptive context.** FastWAM learns from video prediction while supporting actions without explicit future imagination (Yuan et al., 2026). WAM-TTT adapts video-side memory, whereas RoboTTT maintains recurrent fast weights on the action side (Feng et al., 2026; Jiang et al., 2026). These approaches differ in where demonstration information enters prediction and whether inference updates memory. WildWAM learns its reference interface through Internet pre-adaptation and retains fixed parameters at inference. Figure 2 compares the information paths.

**Retrieval and human–robot data.** FlowRetrieval and STRAP retrieve robot experience using motion and subsequence similarity (Lin et al., 2025; Memmel et al., 2025); HAND retrieves robot behaviors from demonstrated hand paths, while RAM transfers retrieved affordances (Hong et al., 2026; Kuang et al., 2025). R+X searches long human videos; SiMDex uses robot queries, and RoboTok uses human-query motion representations (Papagiannis et al., 2025; Lin et al., 2026; Qian et al., 2026). Such matching supplies behavioral candidates; transferable procedures also require temporal extent, object roles, and completion evidence. WildWAM connects EgoDex, EPIC-KITCHENS, and Open-AoE to WAM learning through event localization and applicability annotation (Hoque et al., 2026; Damen et al., 2022; Li et al., 2026).

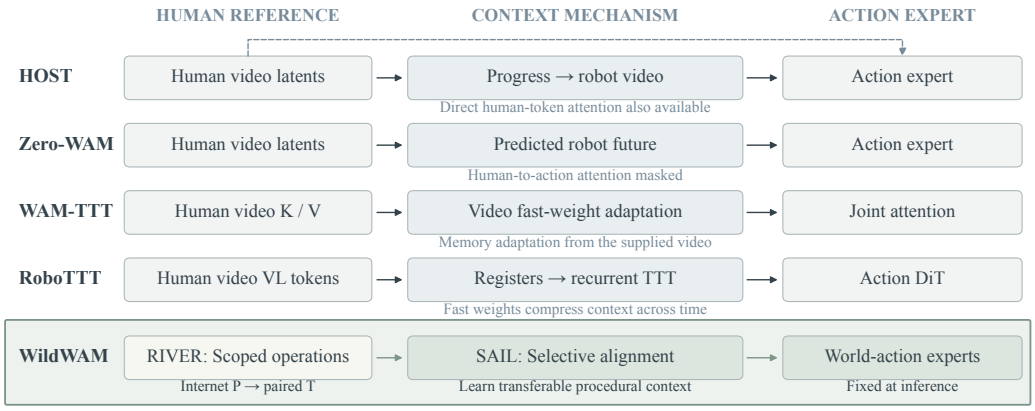



Figure 2: **How a human video reaches robot actions.** HOST (Chen et al., 2026), Zero-WAM (Zhou et al., 2026a), WAM-TTT (Feng et al., 2026), and RoboTTT (Jiang et al., 2026) use different prediction and memory interfaces. WildWAM couples RIVER-retrieved operations with SAIL's direct visual context and history-conditioned reading.

# 3 WILDWAM

Let $R_i = \{(o_t, s_t, a_t)\}_{t=1}^{T_i}$ be a robot episode with instruction $\ell_i$ and embodiment $b_i$, and let $h = (v, u, w)$ denote an exact interval $[u, w)$ of a human source video. WildWAM combines RIVER to retrieve annotated correspondences $(R_i, h)$ with SAIL to learn a reference-conditioned WAM. Figure 1 shows the retrieval and alignment interfaces; Figure 2 places the conditioning mechanism alongside recent approaches.

## 3.1 RIVER: IN-THE-WILD EVENT RETRIEVAL

RIVER searches complementary visual embeddings, task text, and source descriptions. Their candidate sets are merged by exact source–interval identity; WeMM supplies multimodal reranking (Zhou et al., 2026b). The production index contains 715,565 base intervals from three human corpora. Long-video candidates retain source times so adjacent context can be recovered and re-encoded.

**Event-aware temporal selection.** For a nominated seed, let $\mathcal{W}(e)$ contain eligible windows of the same event, including the seed itself. Native action intervals $\mathcal{E}_i = \{e_1, \ldots, e_J\}$ define the required stages, with union $U_i = \bigcup_{e \in \mathcal{E}_i} e$. We measure minimum stage coverage, total recall, and temporal precision:

$$c(w) = \min_{e \in \mathcal{E}_i} \frac{|w \cap e|}{|e|}, \qquad r(w) = \frac{|w \cap U_i|}{|U_i|}, \qquad p(w) = \frac{|w \cap U_i|}{|w|}. \tag{1}$$

Minimum coverage protects every required stage; the harmonic mean of recall and precision favors complete, concentrated windows. We select lexicographically:

$$w^* = \operatorname*{lexmax}_{w \in \mathcal{W}(e)} \left( c(w), \frac{2r(w)p(w)}{r(w) + p(w)}, \mathbf{1}[w = w_{\text{seed}}], -|w| \right). \tag{2}$$

The harmonic term is zero for zero overlap. Gaps between annotations do not count as action time. After coverage and concentration, ties favor the seed, then shorter windows, then stable IDs. Explicit multi-operation queries may include adjacent actions; videos without atomic boundaries retain their original interval. Selection uses interval metadata before ordered visual verification.

**Role and transition verification.** A sequence verifier evaluates manipulated-object function, receiver role, direction, phase order, final state, and required count or chain. Contradictions demote candidates; supported completion raises their evidence tier, with embedding rank breaking ties. Source quotas prevent repeated intervals from dominating a reference list. The controlled retrieval

study holds candidate and sampled-frame budgets fixed within each scope. The three-corpus production pipeline additionally uses per-corpus nominations and context proposals to cover operations beyond the original benchmark.

### 3.2 Annotate the scope of correspondence

Independent robot and human observations record visible operations, initial/final states, and event evidence. Pair annotation then assigns $y_{ih} \in \{\text{full}, \text{partial}, \text{none}, \text{unknown}\}$, with separate ordinal functional and execution-compatibility scores. A partial reference can support one robot event without representing the full program. Event links retain their asset IDs, intervals, and evidence rather than imposing frame synchronization.

During Internet pre-adaptation (P), full references condition robot action/video training; all known pair categories contribute representation and relation supervision. Unpaired robot episodes retain their robot objective. During paired adaptation (T), references are sampled from the collected same-split admissible pool. P uses the robot instruction; T uses the human demonstration and causal robot history without external task text. This separates scalable Internet supervision from the collected reference distribution.

### 3.3 SAIL: Selective in-context alignment

SAIL reads human context conditioned on causal robot history. A frozen video encoder produces $16 \times 4 \times 4$ human features. A trainable reference encoder maps them to $d = 256$ dimensions; spatial pooling preserves 16 temporal positions and yields $V_h \in \mathbb{R}^{64 \times d}$. At decision time $t$, robot history tokens encode visual features, states, and past actions, with time encodings. A causal temporal encoder summarizes history as $q_t$. Normalized history conditions eight role queries:

$$Q_t = 32Q_0 + \mathbf{1}_8 \,\mathrm{LN}(q_t)^\top, \qquad \Pi_t^{(k)} = \mathrm{softmax}\left(\frac{(Q_t W_Q^{(k)})(M_h W_K^{(k)})^\top}{\sqrt{d_h}}\right). \tag{3}$$

The memory $M_h = [V_h; v_\varnothing]$ includes a learned null item. Four heads ($d_h = 64$) read reference content conditioned on causal history, yielding residual role tokens:

$$S_t = \mathrm{LN}\Big(Q_t + \mathrm{Concat}_{k=1}^{4}[\Pi_t^{(k)} M_h W_V^{(k)}] W_O\Big). \tag{4}$$

Role-averaged attention over upcoming and next-useful queries gives a normalized distribution $\alpha_{t,j}$ over the 16 human time positions. A soft gate modulates every spatial token at position $j$:

$$g_{t,j} = \max(16\alpha_{t,j}, 0.05)^{\frac{1}{2}\,\mathrm{sigmoid}(\eta)}, \qquad \widetilde{V}_{h,j,p} = g_{t,j} V_{h,j,p}. \tag{5}$$

The positive gate preserves direct visual content while emphasizing relevant phases. Projected history, gated reference, and role readout form the expert context:

$$C_t^a = [C_t^{\mathrm{hist}};\ \widetilde{V}_h W_c;\ \tanh(\gamma) S_t W_s]. \tag{6}$$

The scale $\tanh(\gamma)$ controls the residual readout contribution. Upcoming role features additionally condition the action expert through a learned plan projection. In P, instruction features also enter the reader query and conditioning sequence. A separate four-query reader, whose memory includes a learned null token, predicts future visual/state targets and supplies goal tokens to the video expert. When no reference is supplied, a learned null representation replaces the human visual context. Appendix E gives tensor dimensions and branch-specific inputs.

### 3.4 Learning transferable content and robot dynamics

For robot or human asset $z$, semantic heads align action-summary, initial-state, and final-state representations to frozen text targets. Pair heads predict the known scope class and known functional/compatibility scores:

$$\mathcal{L}_{\mathrm{human}} = \mathcal{L}_{\mathrm{sem}}^h + \tfrac{1}{2}\,\mathrm{CE}(\hat{y}, y) + 0.1 \frac{\sum_j m_j (\hat{s}_j - s_j)^2}{\sum_j m_j}. \tag{7}$$

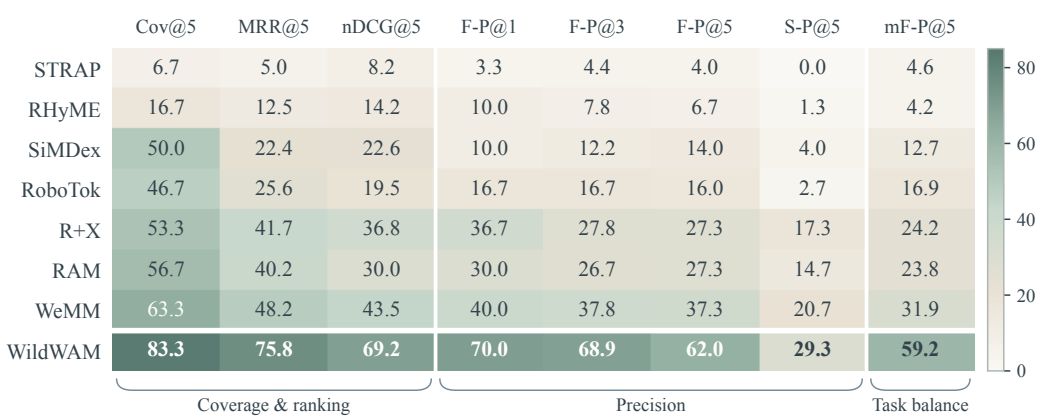



Figure 3: **Retrieving transferable human operations.** Union-scope scores ($\times 100$; higher is better), sorted by F-P@5 ascending. F/S-P: functional/strict precision; Cov: functional coverage; MRR: functional reciprocal rank; nDCG: graded ranking with a common reviewed pool; mF-P: functional precision macro-averaged over task families. Rows denote local retrieval implementations; Appendix D details their inputs, adaptations, and candidate pools.

Unknown fields have $m_j = 0$; objectives with no valid target are omitted. Full-reference control and partial-pair representation learning therefore use different supervision scopes.

The WAM learns joint action/video flow matching (Lipman et al., 2023). Under the implementation convention $x_\sigma = (1-\sigma)y + \sigma\epsilon$, velocity predicts $\epsilon - y$, and the clean estimate is $\hat{y} = x_\sigma - \sigma v_\theta(x_\sigma, \sigma, C_t)$. In addition to the masked flow objectives, we supervise recorded changes:

$$\mathcal{L}_{\Delta a} = \mathrm{SmoothL1}_m(\Delta\hat{A} - \Delta A), \qquad \mathcal{L}_{\Delta v} = \frac{\sum_{\tau,p} m_{\tau p}\omega_{\tau p}\|\hat{Y}_{\tau p} - Y_{\tau p}\|^2}{\sum_{\tau,p} m_{\tau p}\omega_{\tau p}}. \tag{8}$$

Here $\Delta$ denotes adjacent action differences. The latent term emphasizes regions that change relative to the initial observation, using a bounded weight:

$$\omega_{\tau p} = 1 + \min\left(3, \frac{|Y_{\tau p} - Y_{0p}|}{\mathrm{mean}\,|Y - Y_0| + \varepsilon}\right). \tag{9}$$

Channel-averaged latent change defines $\omega$ and is detached from the gradient. Future-goal losses supervise the reference reader and first-frame backbone at an event endpoint and 4/8/16-second horizons, with separate availability masks. The combined control objective is

$$\mathcal{L}_{\mathrm{control}} = \mathcal{L}^a_{\mathrm{FM}} + \mathcal{L}^v_{\mathrm{FM}} + 0.2\mathcal{L}_{\Delta a} + 0.1\mathcal{L}_{\Delta v} + 0.1(\mathcal{L}^{\mathrm{reader}}_{\mathrm{goal}} + \mathcal{L}^{\mathrm{main}}_{\mathrm{goal}} + \mathcal{L}_{\mathrm{event}}). \tag{10}$$

Future observations enter targets, while conditioning contains only the current frame, causal history, and human reference. Coarse P adds $0.1(\mathcal{L}^r_{\mathrm{sem}} + \mathcal{L}_{\mathrm{human}})$; T retains the control objective. Fine P additionally supervises event correspondence and causal progress, enabling reference-reader endpoint targets only when the human video covers that event (Appendix E). Rank-16 cross-attention LoRA (Hu et al., 2022) and the conditioning modules adapt the pretrained backbone; parameters stay fixed at deployment.

## 4 Experiments

The experiments connect three questions: whether retrieval finds transferable operations, whether broader Internet sources add missing task coverage, and whether Internet pairing improves subsequent robot learning.

### 4.1 Retrieval quality and component contributions

The controlled study uses robot queries from 15 collection tasks and 13 functional families. We evaluate 84 configurations in each of four frozen candidate scopes, totaling 336 comparisons. The largest

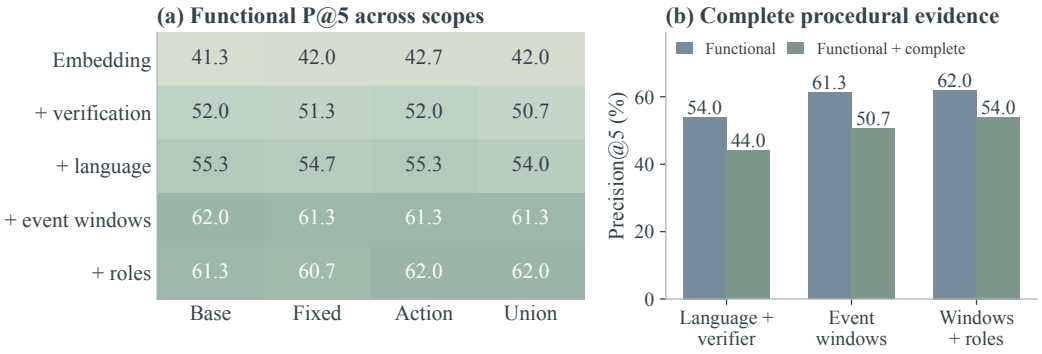



Figure 4: **Event selection works across candidate scopes.** (a) Functional P@5 under four frozen scopes. (b) Functional precision and the jointly confirmed functional-plus-complete-process precision. All within-scope event ablations retain the verifier-slot and temporal-frame budgets.

scope contains 910,373 windows from EgoDex and EPIC-KITCHENS. Functional P@5 measures functional matches, strict P@5 additionally requires strict correspondence, and coverage measures queries with at least one functional hit. Separately reviewed process labels identify complete operations. Appendix D gives the exact rubric, annotation provenance, budgets, and full results.

**Procedural evidence improves retrieval.** WildWAM reaches **70.0% functional P@1, 62.0% functional P@5, and 83.3% coverage** (Figure 3). On the external shared pool, WeMM (Zhou et al., 2026b) yields 37.3% functional P@5, R+X-style (Papagiannis et al., 2025) and RAM (Kuang et al., 2025) yield 27.3%, and the SiMDex-initial (Lin et al., 2026) and RoboTok-local (Qian et al., 2026) proxies yield 14.0% and 16.0%. The figure also includes STRAP (Memmel et al., 2025) and RHyME (Kedia et al., 2025) adaptations. The full pipeline adds event-level temporal selection and ordered role evidence to semantic matching.

**The gain comes from selecting the operation's temporal extent.** Starting from language-augmented retrieval and verification, event-aware selection raises functional P@5 from 54.0% to 61.3% on Union. The gain repeats across all scopes: +6.67, +6.67, +6.00, and +7.33 percentage points for Base, Fixed, Action, and Union. Adding explicit roles yields 62.0% functional precision and 54.0% joint functional-and-complete precision, versus 44.0% for the language/verifier baseline. Figure 4 separates temporal availability from final procedural selection.

### 4.2 Internet diversity supplies missing demonstrations

The production index combines EgoDex (Hoque et al., 2026), EPIC-KITCHENS (Damen et al., 2022), and Open-AoE (Li et al., 2026), accounting for 3.3%, 10.7%, and 86.0% of indexed human intervals, respectively. Across all robot episodes, annotation assigns 42.7% of pair links to full correspondence, 46.3% to partial correspondence, 8.7% to no correspondence, and 2.4% to unknown. Full references cover 80.8% of these robot episodes. Figure 5 shows high-quality examples; Appendix C provides an extended gallery and source-level accounting.

**Open-AoE expands retrieved demonstration diversity and task coverage.** Across all 59,589 G1 episodes, Open-AoE supplies **51.3% of top-1 candidates, 51.7% of top-5 candidates, and 66.6% of distinct retrieved human clips** under four-route fusion (Figure 6). Its top-5 share remains 52.9% when tasks receive equal weight, showing that its contribution extends beyond frequently repeated robot tasks. Appendix C analyzes the source distribution in more detail. In the annotated collection, Open-AoE supports 54.7% of robot episodes with full references, spanning pouring, wiping, folding, grasping, drawer manipulation, and laptop closure.

For examlple, laptop closure illustrates the complementary procedural coverage. Compared with EgoDex/EPIC candidates which show already-closed laptops, cable manipulation, box lids, or oven doors, Open-AoE supplies the required open-to-closed hinged-screen transition for Inspire and Dex3

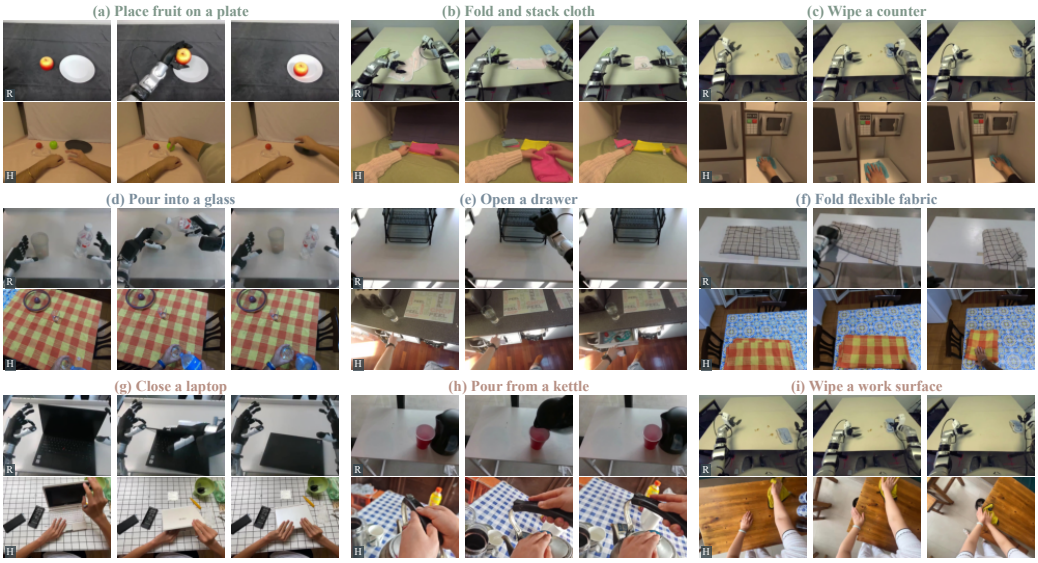



Figure 5: **Robot queries and retrieved human procedures.** EgoDex (a–c), EPIC-KITCHENS (d–f), and Open-AoE (g–i). Each pair shows 3 ordered robot (R, top) and human (H, bottom) frames.

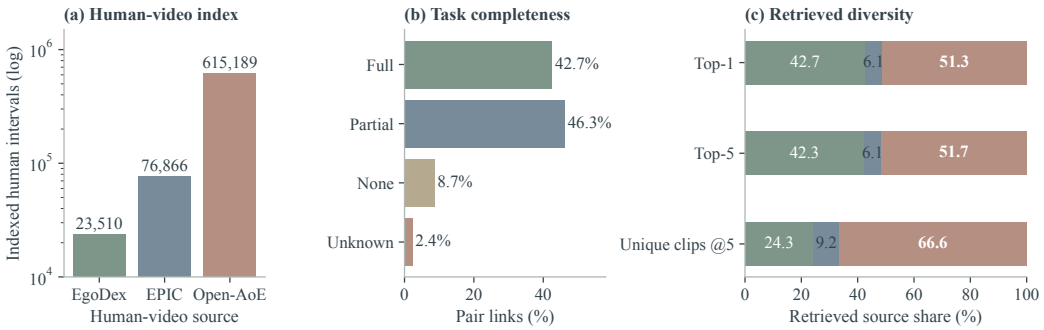



Figure 6: **Scale and source complementarity.** (a) Three-corpus index. (b) Annotated task completeness. (c) Source shares under four-route reciprocal rank fusion across 59,589 robot episodes: top-ranked candidates and distinct human clips retrieved at top five. Colors in (a,c): EgoDex, EPIC-KITCHENS, and Open-AoE; Open-AoE supplies 66.6% of distinct retrieved clips.

episodes without a full EgoDex/EPIC match in the annotated collection. Figure 5 places this example alongside the broader range of retrieved procedures.

### 4.3 A PRETRAINED FOUNDATION ACROSS SIX G1 HAND LAYOUTS

We unify 59,589 robot episodes ($\approx$ 35M frames) across six hand layouts and train FastWAM for 350,000 updates ($\approx$ 1,300 H20 GPU hours). The fixed evaluation panel contains 240 windows from 120 episodes, balanced across layouts, with fixed noise and ten denoising steps. From 1k to 350k updates, macro normalized action MAE falls from 0.16908 to 0.04351 (**74.27%** reduction), while future-video SSIM rises from 0.6280 to 0.6959. All six layouts improve (Appendix B, Figure 10).

### 4.4 INTERNET PRE-ADAPTATION IMPROVES ROBOT ACTION LEARNING

We evaluate target-scene robot episodes with start and middle windows. Stage P uses Internet human-robot links; T uses additional 4,374 collected strong human-robot relations without external task text. Models share the A350,000 initialization, rank-16 LoRA, reference interface, and

Table 1: **Stage-wise real-robot rollout success.** Task cells show cumulative successes at ordered milestones (15 trials each; final count bold), shaded by mean stage progress. Overall shows first-stage / second-stage / final attainment (%, rounded). Pretrained FastWAM uses the foundation only; Internet-only adds Internet pre-adaptation (P); Target-only adds collected-pair adaptation (T); WildWAM uses P followed by T. Appendix F gives stage definitions and later-stage totals.

| Policy | Single-arm manipulation | | Bimanual coordination | | | Overall |
|---|---|---|---|---|---|---|
| | Pick-place | Stack-two | Dual-place | Close-drawer | Replace-press | |
| Pretrained FastWAM | 9/5/**3** | 5/2/**0** | 2/**1** | 2/1/**0** | 3/0/0/**0** | 28/12/**5** |
| Internet-only | 2/1/**0** | 2/0/**0** | 1/**0** | 2/0/**0** | 1/0/0/**0** | 11/1/**0** |
| Target-only | 14/13/**10** | 12/8/**3** | 9/**6** | 12/5/**2** | 10/3/0/**0** | 76/47/**28** |
| **WildWAM (Ours)** | 15/15/**12** | 14/11/**6** | 14/**9** | 12/10/**4** | 13/6/4/**2** | 91/68/**44** |

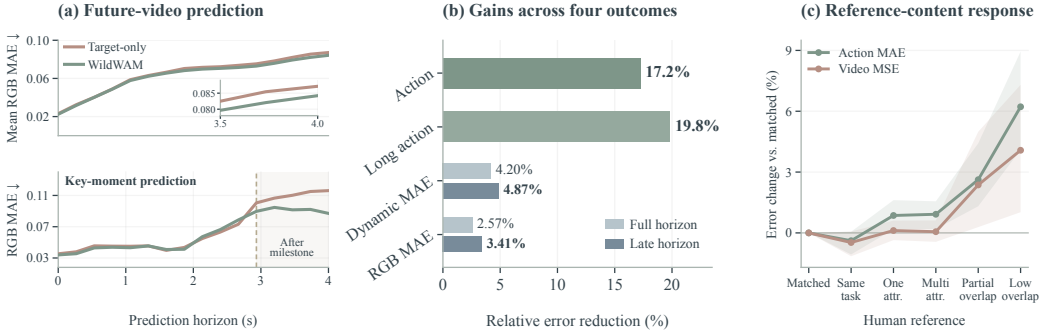



Figure 7: **Internet pairs improve robot prediction.** (a) Mean and key-moment (defined in Appendix F) video errors after matched target-scene adaptation; the inset enlarges the mean curve at 3.5–4.0 s. The vertical line marks a task milestone. (b) Error reductions at a matched target-scene budget; video bars compare full and late horizons. (c) Reference-content interventions: action MAE and dynamic video MSE changes relative to matched references.

T pool. T uses 64k updates, with 64k additional P updates for Internet models (≈ 100 H20 GPU hours). Evaluation uses four denoising steps and predicts 128 actions with 17 future frames (next 4s prediction).

**Internet pre-adaptation improves downstream adaptation.** Table 1 compares no adaptation (Pretrained FastWAM), P alone (Internet-only), T alone (Target-only), and P→T (WildWAM) from a common pretrained foundation. T and P→T share the target-scene pool and T-update budget. Adding Internet pre-adaptation improves all five tasks, raising full-task success from 28.0% to **44.0%** and mean stage progress from 49.4% to **67.7%**. These gains establish the downstream value of retrieved Internet human–robot pairs. Appendix F defines the task milestones.

**Internet pairs deliver consistent downstream prediction gains.** Internet-pair pre-adaptation reduces mean action MAE from 0.15053 to **0.12459**, a **17.2%** relative reduction; long-segment action MAE improves by **19.8%**. With robot context fixed and no task text, same-task reference substitutions leave errors nearly unchanged; low procedure overlap raises action MAE by **6.22%** and dynamic video MSE by **4.08%** (Figure 7c). This intervention shows that both prediction heads respond to human procedural content.

**Internet pairs improve long-horizon video prediction.** Internet pairs reduce full-horizon RGB and dynamic-region MAE by 2.57% and 4.20%; late-horizon reductions reach 3.41% and 4.87% (Figure 7a). The bread-placement example highlights the advantage around a task milestone: final-frame RGB and dynamic-region errors decrease by **25.05%** and **35.15%**. Figure 14, provides the complete temporal curves and corresponding ground-truth, Target-only, and WildWAM frames before and after this transition.

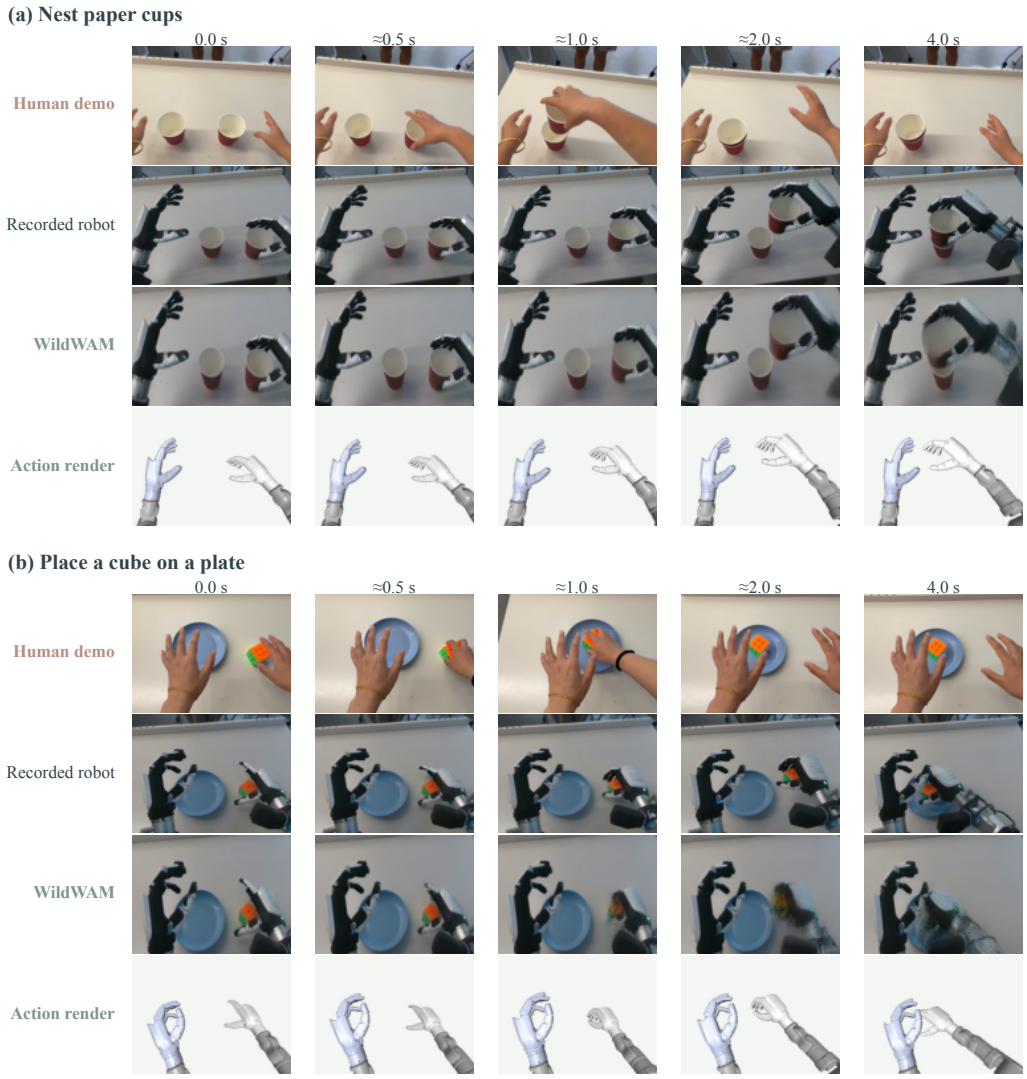



Figure 8: **In-context manipulation with human demonstration.** We provide time-aligned human references, ground truth robot frames (excluded from the training set), predicted futures, and predicted actions rendered with the Unitree G1 + Inspire Hand model. Human frames show demonstration progress; robot rows share the displayed horizon. Additional tasks appear in Figure 15.

## 5 Conclusion

WildWAM combines in-the-wild retreival (RIVER) and procedural alignment (SAIL), which turns unstructured Internet video into usable robot context by coupling event-aware retrieval with selective alignment in a pretrained world-action model. Temporal windows and role/outcome verification recover transferable procedures from long human recordings. Across three human-video sources, broader in-the-wild coverage supplies distinctive references; Internet-pair pre-adaptation then improves held-out robot action and future-video prediction. Together with the unified G1 corpus and traceable pair collection, these results establish a route from human observation to robot learning.

Our evaluation focuses on egocentric human and robot videos, where hand–object progression is visible in both domains. Extending procedural matching and alignment to mixed and third-person viewpoints could enlarge the pool of usable everyday demonstrations. Building visual feature caches and detailed event labels has an offline cost; reusable representations, adaptive evidence sampling, and efficient verifiers can amortize it as the library grows.

## Disclosure of generative AI use

In this work, we used large language models to polish manuscript language, assist with organizing experimental records, summarizing results, and framework diagram guidance. These summaries were checked against the underlying experiment outputs and data records. We also used GPT-5.6 Sol (`gpt-5.6-sol`) to automatically annotate human–robot video pairs, including coarse task-level correspondence and fine event-level alignment; Appendix G describes the annotation schema and its use in training. The authors reviewed the AI-assisted manuscript and take responsibility for the final text, claims, results, and artifacts.

## References


Guangyan Chen, Meiling Wang, Te Cui, Zichen Zhou, Qi Shao, Shalfun Li, Hang Su, Roy Gan, Hao Wang, Mengyin Fu, Yi Yang, and Yufeng Yue. HOST: Robots acquire manipulation skills in seconds from a single human video. *arXiv preprint arXiv:2607.20033*, 2026.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling Egocentric Vision: Collection, Pipeline and Challenges for EPIC-KITCHENS-100. *International Journal of Computer Vision*, 130(1):33–55, 2022. doi: 10.1007/s11263-021-01531-2.

Mikita Dvornik, Isma Hadji, Konstantinos G. Derpanis, Animesh Garg, and Allan D. Jepson. Drop-DTW: Aligning Common Signal Between Sequences While Dropping Outliers. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Yusen Feng, Bingchen Han, Jiangran Lyu, Kai Liu, Yixin Zheng, Yuxuan Wan, Weiheng Liu, Sun Han, Ruiqin Li, Yulong Zhang, Fangfu Liu, Xuesong Shi, Libin Liu, Yizhou Wang, Zhizheng Zhang, and He Wang. WAM-TTT: Steering world-action models by watching human play at test time. *arXiv preprint arXiv:2607.06988*, 2026.

Liang Heng, Yihe Tang, Jiajun Xu, Henghui Bao, Di Huang, and Yue Wang. HumDex: Humanoid Dexterous Manipulation Made Easy. *arXiv preprint arXiv:2603.12260*, 2026.

Matthew Hong, Anthony Liang, Kevin Kim, Harshitha Rajaprakash, Jesse Thomason, Erdem Bıyık, and Jesse Zhang. HAND Me the Data: Fast Robot Adaptation via Hand Path Retrieval. In *IEEE International Conference on Robotics and Automation*, 2026.

Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. EgoDex: Learning dexterous manipulation from large-scale egocentric video. In *International Conference on Learning Representations*, 2026.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022.

Thomas Hummel, Shyamgopal Karthik, Mariana-Iuliana Georgescu, and Zeynep Akata. EgoCVR: An Egocentric Benchmark for Fine-Grained Composed Video Retrieval. In *European Conference on Computer Vision*, 2024.

Vidhi Jain, Maria Attarian, Nikhil J. Joshi, Ayzaan Wahid, Danny Driess, Quan Vuong, Pannag R. Sanketi, Pierre Sermanet, Stefan Welker, Christine Chan, Igor Gilitschenski, Yonatan Bisk, and Debidatta Dwibedi. Vid2Robot: End-to-end video-conditioned policy learning with cross-attention transformers. In *Robotics: Science and Systems*, 2024.

Yunfan Jiang, Yevgen Chebotar, Ruijie Zheng, Fengyuan Hu, Yunhao Ge, Jimmy Wu, Tianyuan Dai, Scott Reed, Fei-Fei Li, Yuke Zhu, and Linxi Fan. RoboTTT: Context scaling for robot policies. *arXiv preprint arXiv:2607.15275*, 2026.

Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker3: Simpler and Better Point Tracking by Pseudo-Labelling Real Videos. In *IEEE/CVF International Conference on Computer Vision*, pp. 6013–6022, 2025.

Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. EgoMimic: Scaling imitation learning via egocentric video. In *IEEE International Conference on Robotics and Automation*, 2025.

Kushal Kedia, Prithwish Dan, Angela Chao, Maximus Adrian Pace, and Sanjiban Choudhury. One-shot imitation under mismatched execution. In *IEEE International Conference on Robotics and Automation*, 2025.

Yuxuan Kuang, Junjie Ye, Haoran Geng, Jiageng Mao, Congyue Deng, Leonidas Guibas, He Wang, and Yue Wang. RAM: Retrieval-based affordance transfer for generalizable zero-shot robotic manipulation. In *Proceedings of The 8th Conference on Robot Learning*, volume 270, pp. 547–565, 2025.

Robin Kühn, Moritz Schappler, Thomas Seel, and Dennis Bank. Active Stereo-Camera Outperforms Multi-Sensor Setup in ACT Imitation Learning for Humanoid Manipulation. *arXiv preprint arXiv:2603.28422*, 2026.

Zishuo Li, Bowen Yang, Changtao Miao, Kai Zhu, Hao Chen, Qingze Guan, Zhengxing Wu, Wanke Zhan, Yang Sun, Zhiyi Huang, Zitong Shan, Zhenchao Jin, Jiadong Hong, Taowen Wang, Yushi Feng, You Liu, Yibo Wang, Yifan Yang, Zhaowen Zhou, Man Luo, Hao Cheng, Bo Zhang, Jianshu Li, Jiansheng Cai, Guocai Yao, Jize Zhang, Chenhao Lin, Renjing Xu, Lequan Yu, Chao Shen, Chunhua Shen, and Zhe Li. Open-AoE: An open egocentric manipulation dataset and toolchain for embodied learning. *arXiv preprint arXiv:2607.14183*, 2026.

Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. UniVTG: Towards Unified Video-Language Temporal Grounding. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Li-Heng Lin, Yuchen Cui, Amber Xie, Tianyu Hua, and Dorsa Sadigh. FlowRetrieval: Flow-guided data retrieval for few-shot imitation learning. In *Proceedings of The 8th Conference on Robot Learning*, volume 270, pp. 4084–4099, 2025.

Nie Lin, Takehiko Ohkawa, Sijin Chen, Ruoshi Wen, Zhuohang Li, Liqun Huang, Zhengming Zhu, Yiming Bao, Yunfei Li, Minjie Cai, Xiao Ma, Wei Xu, and Yoichi Sato. SiMDex: Mining similar egocentric videos for cross-embodiment dexterous manipulation. *arXiv preprint arXiv:2608.04196*, 2026.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow Matching for Generative Modeling. In *International Conference on Learning Representations*, 2023.

Jiacheng Liu, Pengxiang Ding, Qihang Zhou, Yuxuan Wu, Da Huang, Zimian Peng, Wei Xiao, Weinan Zhang, Lixin Yang, Cewu Lu, and Donglin Wang. TrajBooster: Boosting Humanoid Whole-Body Manipulation via Trajectory-Centric Learning. In *IEEE International Conference on Robotics and Automation*, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019.

Marius Memmel, Jacob Berg, Bingqing Chen, Abhishek Gupta, and Jonathan Francis. STRAP: Robot sub-trajectory retrieval for augmented policy learning. In *International Conference on Learning Representations*, 2025.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-JEPA 2.1: Unlocking Dense Features in Video Self-Supervised Learning. In *European Conference on Computer Vision*, 2026.

NVIDIA. Cosmos-Embed1. Official model documentation, 2025. URL https://docs.nvidia.com/nim/cosmos-embed1/1.0.0/introduction.html.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*, 2024.

Georgios Papagiannis, Norman Di Palo, Pietro Vitiello, and Edward Johns. R+X: Retrieval and execution from everyday human videos. In *IEEE International Conference on Robotics and Automation*, 2025.

Baoqi Pei, Guo Chen, Jilan Xu, Yuping He, Yicheng Liu, Kanghua Pan, Yifei Huang, Yali Wang, Tong Lu, Limin Wang, and Yu Qiao. EgoVideo: Exploring Egocentric Foundation Model and Downstream Adaptation. *arXiv preprint arXiv:2406.18070*, 2024.

Howard Qian, Yiting Chen, Yunfei Xie, Kejia Ren, Podshara Chanrungmaneekul, Gaotian Wang, Bowen Wen, Chen Wei, and Kaiyu Hang. RoboTok: An internet-scale data engine for human demonstration retrieval and dexterous manipulation learning. *arXiv preprint arXiv:2609.03199*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Rutav Shah, Shuijing Liu, Qi Wang, Zhenyu Jiang, Sateesh Kumar, Mingyo Seo, Roberto Martín-Martín, and Yuke Zhu. MimicDroid: In-context learning for humanoid robot manipulation from human play videos. In *IEEE International Conference on Robotics and Automation*, 2026.

Modi Shi, Shijia Peng, Jin Chen, Haoran Jiang, Tianyu Li, Di Huang, Ping Luo, Hongyang Li, and Li Chen. EgoHumanoid: Unlocking In-the-Wild Loco-Manipulation with Robot-Free Egocentric Demonstration. In *Robotics: Science and Systems*, 2026.

Apoorv Vyas, Heng-Jui Chang, Cheng-Fu Yang, Po-Yao Huang, Luya Gao, Julius Richter, Sanyuan Chen, Matthew Le, Piotr Dollár, Christoph Feichtenhofer, Ann Lee, and Wei-Ning Hsu. Pushing the Frontier of Audiovisual Perception with Large-Scale Multimodal Correspondence Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30172–30182, 2026.

Songlin Wei, Hongyi Jing, Boqian Li, Zhenyu Zhao, Jiageng Mao, Zhenhao Ni, Sicheng He, Jie Liu, Xiawei Liu, Kaidi Kang, Sheng Zang, Weiduo Yuan, Marco Pavone, Di Huang, and Yue Wang. $\Psi_0$: An Open Foundation Model Towards Universal Humanoid Loco-Manipulation. In *Robotics: Science and Systems*, 2026.

Shihan Wu, Xuecheng Liu, Shaoxuan Xie, Pengwei Wang, Xinghang Li, Bowen Yang, Zhe Li, Kai Zhu, Hongyu Wu, Yiheng Liu, Zhaoye Long, Runtian Xu, Yue Wang, Chong Liu, Dihan Wang, Ziqiang Ni, Xiang Yang, You Liu, Ruoxuan Feng, Lei Zhang, Denghang Huang, Chenghao Jin, Anlan Yin, Xinlong Wang, Zhenguo Sun, Junkai Zhao, Mengfei Du, Mingyu Cao, Xiansheng Chen, Hongyang Cheng, Xiaojie Zhang, Yankai Fu, Ning Chen, Cheng Chi, Sixiang Chen, Huaihai Lyu, Xiaoshuai Hao, Yequan Wang, Bo Lei, Dong Liu, Xi Yang, Yance Jiao, Tengfei Pan, Yunyan Zhang, Songjing Wang, Ziqian Zhang, Xu Liu, Ji Zhang, Caowei Meng, Zhizheng Zhang, Jiyang Gao, Song Wang, Xiaokun Leng, Zhiqiang Xie, Zhenzhen Zhou, Peng Huang, Wu Yang, Yandong Guo, Yichao Zhu, Suibing Zheng, Hao Cheng, Xinmin Ding, Yang Yue, Huanqian Wang, Chi Chen, Jingrui Pang, YuXi Qian, Haoran Geng, Lianli Gao, Haiyuan Li, Bin Fang, Gao Huang, Yaodong Yang, Hao Dong, He Wang, Hang Zhao, Yadong Mu, Di Hu, Hao Zhao, Tiejun Huang, Shanghang Zhang, Yonghua Lin, Zhongyuan Wang, and Guocai Yao. RoboCOIN: An Open-Sourced Bimanual Robotic Data Collection for Integrated Manipulation. *arXiv preprint arXiv:2511.17441*, 2025.

Amber Xie, Rahul Chand, Dorsa Sadigh, and Joey Hejna. Data Retrieval with Importance Weights for Few-Shot Imitation Learning. In *Proceedings of The 9th Conference on Robot Learning*, volume 305, pp. 1–16, 2025.

Bowen Yang, Zishuo Li, Yang Sun, Changtao Miao, Yifan Yang, Man Luo, Xiaotong Yan, Feng Jiang, Jinchuan Shi, Yankai Fu, Ning Chen, Junkai Zhao, Pengwei Wang, Guocai Yao, Shanghang Zhang, Hao Chen, Zhe Li, and Kai Zhu. AoE: Always-on Egocentric Human Video Collection for Embodied AI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 3893–3902, 2026.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-WAM: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*, 2026.

Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning Video Representations from Large Language Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6586–6597, 2023.

Zhenyu Zhao, Hongyi Jing, Xiawei Liu, Jiageng Mao, Abha Jha, Hanwen Yang, Rong Xue, Sergey Zakharov, Vitor Guizilini, and Yue Wang. Humanoid Everyday: A Comprehensive Robotic Dataset for Open-World Humanoid Manipulation. In *Embodied AI Workshop at CVPR*, 2026.

Jiaming Zhou, Qihang Zhang, Gangwei Xu, Cunxin Fan, Yujie Zhao, Ruilin Wang, Yiming Luo, Shuai Yang, Xing Zhu, Yujun Shen, Junwei Liang, and Yinghao Xu. Zero-WAM: In-context world-action modeling from human videos for open-ended task generalization. *arXiv preprint arXiv:2608.26103*, 2026a.

Junjie Zhou, Ke Mei, Lei Li, Tianyi Wang, Fengyun Rao, and Jing Lyu. WeMM-Embedding: Wechat multi-modal embedding technical report. *arXiv preprint arXiv:2608.24053*, 2026b.

## APPENDIX CONTENTS


**A Data collections and temporal contracts** . . . . . **14**
A.1 Unified robot foundation . . . . . 14
A.2 Collected target-scene demonstrations . . . . . 16
**B Foundation training and six-layout prediction** . . . . . **16**
**C Verified pairs and corpus-wide source complementarity** . . . . . **17**
C.1 Three-corpus retrieval and annotation . . . . . 17
C.2 Why Open-AoE expands corpus-wide retrieval . . . . . 17
C.3 Corpus-wide automatic nomination and diversity . . . . . 17
**D Retrieval evaluation and all configurations** . . . . . **22**
D.1 Frozen scopes, metrics, and annotation . . . . . 22
D.2 Within-scope budget controls . . . . . 22
D.3 Complete comparison results . . . . . 23
D.4 Event selection and representation controls . . . . . 25
**E Reference-learning architecture and experimental details** . . . . . **25**
E.1 Reference conditioning and supervision scopes . . . . . 25
E.2 Data, optimization, and evaluation . . . . . 26
E.3 Video prediction over task progress . . . . . 26
**F Real-robot task stages and progress metric** . . . . . **29**
**G Automatic annotation from coarse scope to fine correspondence** . . . . . **30**
G.1 Corpus-scale annotation and independent observations . . . . . 30
G.2 Coarse scope and score semantics . . . . . 30
G.3 Events, programs, and local transfer . . . . . 31
G.4 Task specification, causal prefixes, and future targets . . . . . 32


# A DATA COLLECTIONS AND TEMPORAL CONTRACTS

## A.1 UNIFIED ROBOT FOUNDATION

The unified Unitree G1 foundation contains 59,589 episodes in 290 partition and 270 distinct task descriptions, with 35,383,513 output frames and 327.625 hours at 30 Hz. The interface preserves embodiment identity, coordinate meaning, and source provenance. Table 2 gives the six layouts; Table 3 gives the conversion-route composition.

The robot corpus includes the Unitree G1 + Inspire Hand teleoperation data from AoE (Yang et al., 2026), distinct from the Open-AoE human-video retrieval corpus (Li et al., 2026). Other research-associated sources include EgoHumanoid (Shi et al., 2026), Humanoid Everyday (Zhao et al., 2026), HumDex (Heng et al., 2026), $\Psi_0$ (Wei et al., 2026), RoboCOIN (Wu et al., 2025), TrajBooster (Liu et al., 2026), and the Unified Ablation Framework (Kühn et al., 2026).

Table 2: Unified G1 layouts. Hours count output rows at 30 Hz, including repeated observations after resampling. Hand units are meters for Dex1 and radians otherwise.

| Layout | Dimensions | Partitions | Episodes | Frames | Hours |
|---|---|---|---|---|---|
| brainco | 41 | 12 | 2,798 | 2,704,154 | 25.038 |
| dex1 | 31 | 47 | 36,381 | 20,935,174 | 193.844 |
| dex3 | 43 | 201 | 14,479 | 7,566,261 | 70.058 |
| inspire_dfq | 41 | 5 | 1,499 | 1,294,750 | 11.988 |
| inspire_ftp | 41 | 14 | 3,862 | 2,461,879 | 22.795 |
| wuji | 69 | 11 | 570 | 421,295 | 3.901 |
| Total | — | 290 | 59,589 | 35,383,513 | 327.625 |

The shared 29-coordinate body block orders left arm (0–6), right arm (7–13), left leg (14–19), right leg (20–25), and waist (26–28). Layout-specific hand drivers follow, left before right. BrainCo and both Inspire variants have 41 coordinates but retain distinct model bindings. Driver count represents

Table 3: Composition by conversion route. Route names identify source-processing groups, not semantic task classes.

| Route | Partitions | Episodes | Frames | Hours |
|---|---|---|---|---|
| aoe | 6 | 403 | 270,135 | 2.501 |
| cloudwalk | 2 | 200 | 83,998 | 0.778 |
| egohumanoid | 1 | 50 | 35,804 | 0.332 |
| everyday | 123 | 4,024 | 1,758,722 | 16.284 |
| gr00t_n17 | 1 | 402 | 171,625 | 1.589 |
| gr00t_teleop | 4 | 1,095 | 184,751 | 1.711 |
| humdex | 11 | 570 | 421,295 | 3.901 |
| psi_processed | 8 | 625 | 721,507 | 6.681 |
| psi_raw | 6 | 207 | 135,913 | 1.258 |
| robocoin | 44 | 6,255 | 2,924,296 | 27.077 |
| trajbooster | 1 | 8 | 2,831 | 0.026 |
| uaf | 2 | 261 | 299,251 | 2.771 |
| unibot | 32 | 33,276 | 18,282,246 | 169.280 |
| unitree_brainco | 8 | 1,598 | 1,655,918 | 15.333 |
| unitree_dex1 | 2 | 552 | 562,357 | 5.207 |
| unitree_dex3 | 12 | 2,850 | 2,299,528 | 21.292 |
| vla12 | 12 | 2,399 | 1,966,968 | 18.213 |
| wbt | 15 | 4,814 | 3,606,368 | 33.392 |

the exported control coordinates, including model coupling conventions, rather than the number of rendered meshes.

For source coordinate $j$, the conversion applies a recorded permutation $\pi_b$, scale $\alpha_{b,j}$, and offset $\beta_{b,j}$:

$$q_j^{\text{model}} = \alpha_{b,j} q_{\pi_b(j)}^{\text{source}} + \beta_{b,j}. \tag{11}$$

Action and state are mapped independently with absolute-position semantics. Temporal resampling uses a shared modality-index plan. Fixed body completion is recorded where a source lacks measured body channels; actual leg/waist motion is retained where available. Joint values are not clipped during conversion.

Source member lists, episode metadata, and video boundaries jointly determine episode identity. The median output duration is 16.933 s (quartiles 11.433/25.067 s), the 95th percentile is 41.767 s, and the maximum is 262.867 s. Output hours include repeated observations introduced by resampling. Structural checks cover every partition's episode continuity, row counts, coordinate names, frame rate, and egocentric RGB key.

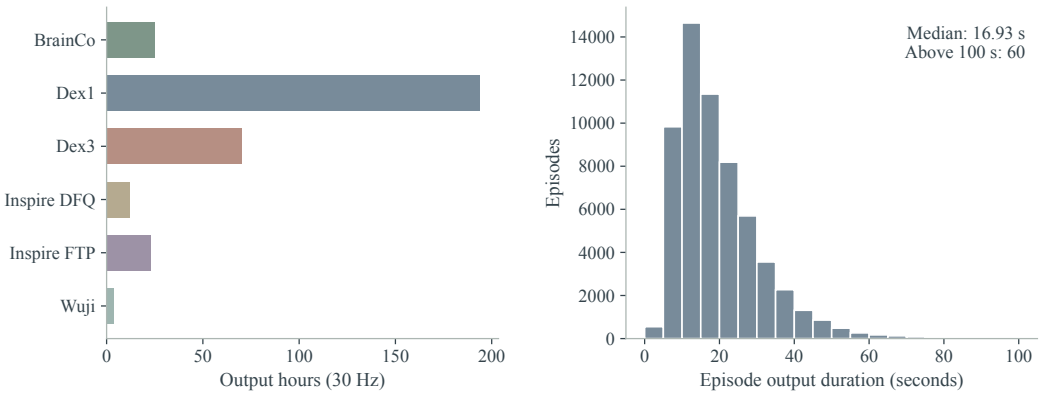



Figure 9: Robot pretraining data composition by layout and episode duration. The histogram displays durations through 100 seconds and records the remaining count in the panel.

### A.2 COLLECTED TARGET-SCENE DEMONSTRATIONS

The collected inventory contains 241 robot episodes ($\approx 1$ hour) and 701 retained human demonstrations. Human demonstrations are organized by task variant into 27 strong-reference categories containing 701 videos. Across the inventory, the full 241 robot episodes admit 6,146 many-to-many strong-reference relations. Pair relations indicate task-compatible demonstrations rather than synchronized human–robot recordings.

The robot stream is 640×480 RGB at 30 Hz with 41-dimensional Inspire FTP action/state. The 15 leg/waist coordinates are fixed standing-pose completion. Arm actions retain recorded pre-rate-limit IK targets, while state retains feedback. Human videos retain native 1920×1080 RGB and presentation timestamps. Cropped versions have explicit source-time mappings.

Stage T uses only collected strong human–robot pairs. After the training/evaluation split, its eligible pool contains 202 robot episodes, 594 distinct human demonstrations, and 4,374 admissible strong-reference relations. Each training example selects a robot window and a human reference from that robot's eligible training pool; inputs comprise the human video and causal robot context, without external task text. Internet references are used exclusively for stage P and are excluded in stage T. Evaluation retains held-out robot episodes with strong references, each providing start and middle windows. Recordings and aliases stay within their assigned split before reference relations are formed.

## B FOUNDATION TRAINING AND SIX-LAYOUT PREDICTION

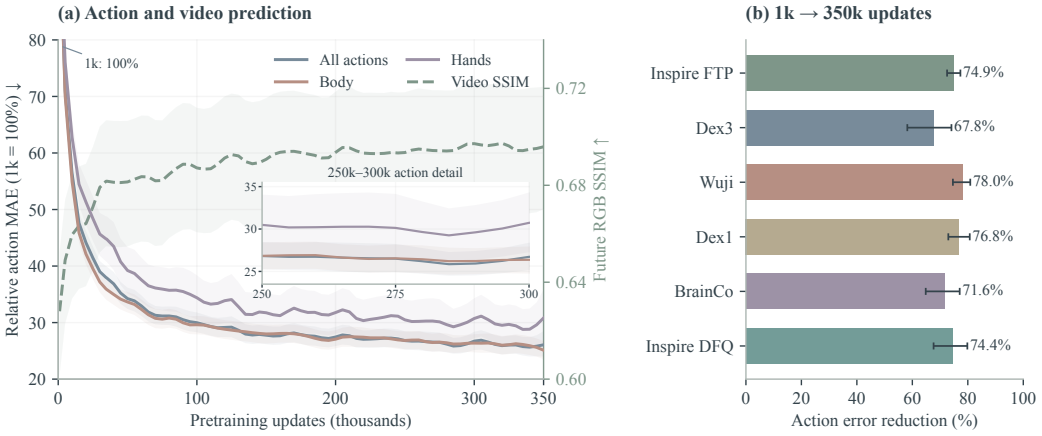



Figure 10: **A common world-action foundation learns all six embodiments.** (a) Action MAE relative to 1k (solid, left) and future-video SSIM (dashed, right), from 1k to 350k. Each action metric is normalized within layout before averaging. The inset expands 100k–350k action trends. Curves use 5k-bandwidth Gaussian smoothing with measured endpoints; shading shows pointwise 95% episode-bootstrap intervals. (b) Per-layout normalized action-MAE reduction over the same interval, with paired 95% intervals.

Training spans 350,000 updates; the latest fixed-panel evaluation is also at 350,000. The panel has 20 episodes per hand layout and two windows per episode, giving 240 windows. Its 72 checkpoints from 1k to 350k use the same episode panel, seed 42, and ten denoising steps. The final training configuration uses eight H20 GPUs, per-device batch 8, no gradient accumulation, and effective batch 64, totaling $\approx$ 1,300 H20 GPU hours.

At 350k, macro normalized action-only MAE across all valid action coordinates is 0.04351 and future RGB SSIM is 0.6959. The paired 1k-to-350k macro action-error reduction is 74.27%, with episode-bootstrap 95% interval [71.86,76.34]. Every layout improves; individual reductions range from 67.76% (Dex3) to 78.03% (Wuji).

Figure 10 separates all-action, body, and hand errors. For each metric, each layout's MAE is divided by its own 1k value before the six ratios are averaged. This preserves native body/hand units within layouts, including Dex1's linear hand coordinate. The absolute all-action MAE reported above instead averages normalized-coordinate errors directly. No value is assigned to step zero.

The last 1,000-update means are 0.166175 total loss, 0.007676 action loss, and 0.158500 video loss. The collected log contains 35,100 finite records at ten-update intervals, and total loss equals the action/video sum.

For uncertainty, both windows of an episode stay together. We resample 20 episode means within each layout 5,000 times under the same seed, reusing episode indices across all checkpoints and the 1k baseline. Relative errors are recomputed in every draw. The Gaussian kernel acts on means and bootstrap trajectories alike, with first and last observations retained; pointwise quantiles yield the shaded intervals. SSIM measures structural similarity of predicted and recorded RGB frames. At 350k, raw predictions exceed a modeled joint limit in 165/240 windows, versus 28/240 for recorded targets; these diagnostics do not enter the action-MAE comparison.

# C Verified pairs and corpus-wide source complementarity

## C.1 Three-corpus retrieval and annotation

The production feature snapshot contains 23,510 EgoDex episodes, 76,866 EPIC atomic clips, and 615,189 Open-AoE atomic clips. It retains 477,252 distinct linked text descriptions. Visual, text-to-video, video-to-description, and lexical routes nominate candidates within each corpus. Per-corpus pools and source quotas preserve diverse evidence before expensive reranking. Context expansion reads source video and re-encodes the resulting interval.

The multi-embodiment annotation release includes 100k annotated links across all 59,589 G1 episodes, contain 42.7% full, 46.3% partial, 8.7% none, and 2.4% unknown relations (All the six robotic hand are included – Inspire FTP, Inspire DFQ, Dex3, Dex1, BrainCo, Wuji). During annotating, each claim stores its value, status, evidence IDs, and uncertainty reasons. Known masks distinguish a scored zero from an unknown field. Local correspondences link robot and human event IDs; event identity is scoped by asset.

## C.2 Why Open-AoE expands corpus-wide retrieval

Across all 59,589 G1 episodes, four-route reciprocal rank fusion retrieves Open-AoE for **51.3% of top-1 positions and 51.7% of top-five positions**, without source quotas. Open-AoE supplies **66.6% of distinct human clips** appearing in the fused top-five results (Figure 6c). Its contribution therefore extends beyond repeated nominations of a few demonstrations: the Internet-scale library provides the largest share of both ranked references and distinct retrieved procedures.

This contribution persists across tasks. Open-AoE appears in the top-five results for **85.6% of task texts** and supplies a majority of their candidates for **54.4%**, leading the three sources on both measures (Figure 12a). Giving each task equal weight yields a 52.9% Open-AoE share, so the result is not driven by frequently repeated robot tasks. Opening, closing, folding, and cleaning procedures illustrate this breadth; the galleries in Figures 5 and 11 show concrete egocentric references, including laptop closure. These corpus-wide measures describe retrieved candidates and source diversity; Section C.3 specifies the full-inventory ranking protocol and source accounting.

## C.3 Corpus-wide automatic nomination and diversity

**Scale and retrieval protocol.** The global nomination stage searches all 59,589 G1 episodes against the frozen inventory of 715,565 human clips. The collection spans 270 task texts, 290 partitions, and six hand configurations. Four routes independently rank the full human inventory: Cosmos video embeddings, WeMM video embeddings, WeMM video with robot task text, and caption BM25. Reciprocal rank fusion combines their global top-100 lists,

$$s_{\mathrm{RRF}}(r,h)=\sum_{m=1}^{4}\frac{\mathbf{1}[h\in\mathcal{C}_m^{100}(r)]}{60+\mathrm{rank}_m(h\mid r)}. \tag{12}$$

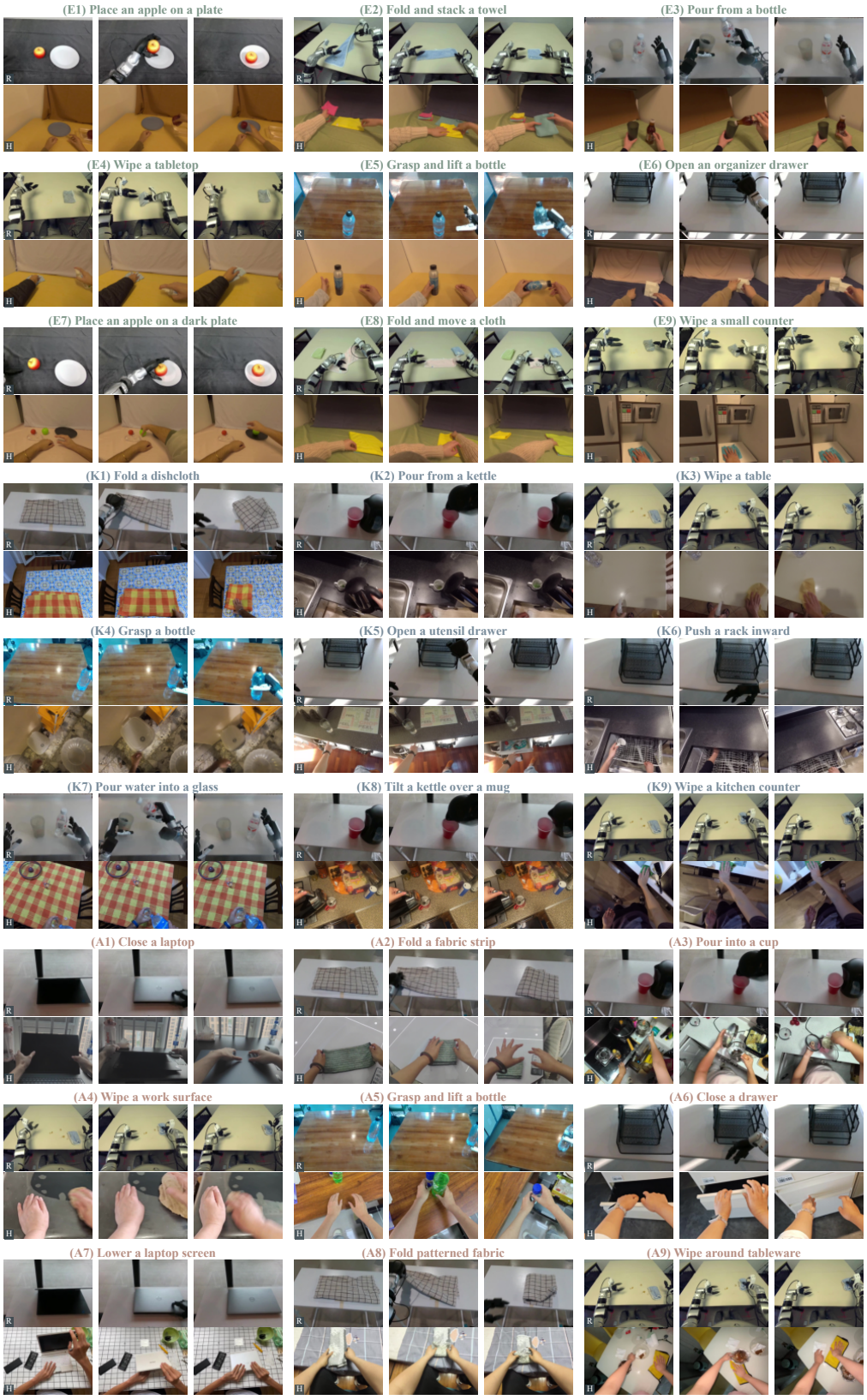



Figure 11: **Additional retrieved human procedures.** Nine full-reference pairs per corpus: EgoDex (E1–E9), EPIC-KITCHENS (K1–K9), and Open-AoE (A1–A9). Each pair shows ordered robot (R, top) and human (H, bottom) frames.

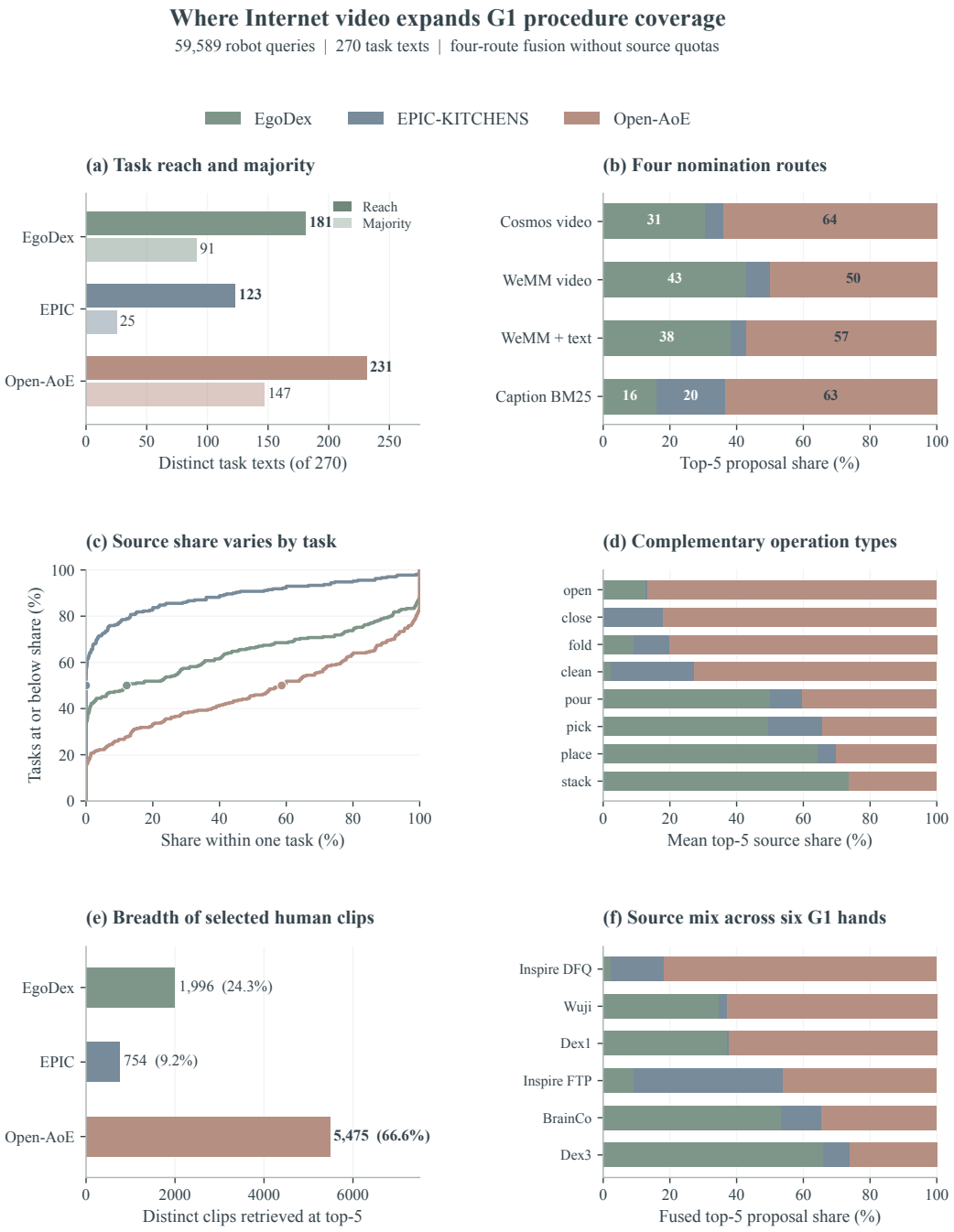



Figure 12: **Internet video broadens task-level retrieval.** (a) Tasks with any top-five proposal and tasks with a source majority. (b) The four nomination routes before fusion. (c) Per-task source-share distributions. (d) Task-weighted source shares for eight operation verbs. (e) Distinct clips appearing in fused top-five results. (f) Fused source mix across six G1 hand layouts.

Fusion uses no dataset quota; separate per-library top-30 diagnostics do not enter its score. The stored index contains 1,191,780 automatic proposals before functional verification. Annotated accuracy and pair coverage are defined in Sections D and C.

**Task reach and route dependence.** Figure 12 shows that Open-AoE contributes top-five candidates to 231 of 270 (over 85%) tasks and supplies a majority of retrieved candidates for some specific tasks, the largest count among the three libraries on both measures. This breadth includes opening, closing, folding, and cleaning operations. The four nomination routes have different source mixtures: caption BM25 gives EPIC-KITCHENS a larger share than the visual routes, while Cosmos video favors Open-AoE. Fusion selects Open-AoE for 51.7% of top-five positions without a source quota.

**Human-source breadth.** Equal weighting over task texts gives Open-AoE 52.9% of fused top-five nominations. The full-corpus top-five draws on 8,225 distinct human clips, of which Open-AoE supplies 66.6%; 35,756 robot queries receive five candidates from distinct source groups. For source-group accounting, EgoDex uses episode identifiers, EPIC-KITCHENS original participant videos, and Open-AoE raw-video identifiers after removing segment suffixes.

Table 4: **Human-source exposure in full-corpus fused top-5.** Each row distinguishes candidate inventory, ranked links, selected clips, and selected source groups.

| Human dataset | Inventory | Links | Clips | Groups |
|---|---|---|---|---|
| EgoDex | 23,510 | 125,971 | 1,996 | 1,996 |
| EPIC-KITCHENS | 76,866 | 18,078 | 754 | 256 |
| Open-AoE | 615,189 | 153,896 | 5,475 | 1,888 |

**Complementary routes and source concentration.** Figure 13 relates fusion inputs to candidate breadth. Cosmos and WeMM video share 0.195 top-five candidates on average, whereas WeMM video and WeMM with task text share 0.939. Multiple routes support 96.76% of fused top-five positions; consensus declines smoothly at deeper ranks. The ten most-used source videos account for 15.63%, 48.35%, and 25.93% of within-library links for EgoDex, EPIC-KITCHENS, and Open-AoE, respectively. At a depth of 20, retrieval exposes 16,094 distinct Open-AoE clips, and 56.1% of robot queries have candidates from at least two libraries; the corresponding top-five breadth is 28.1%.

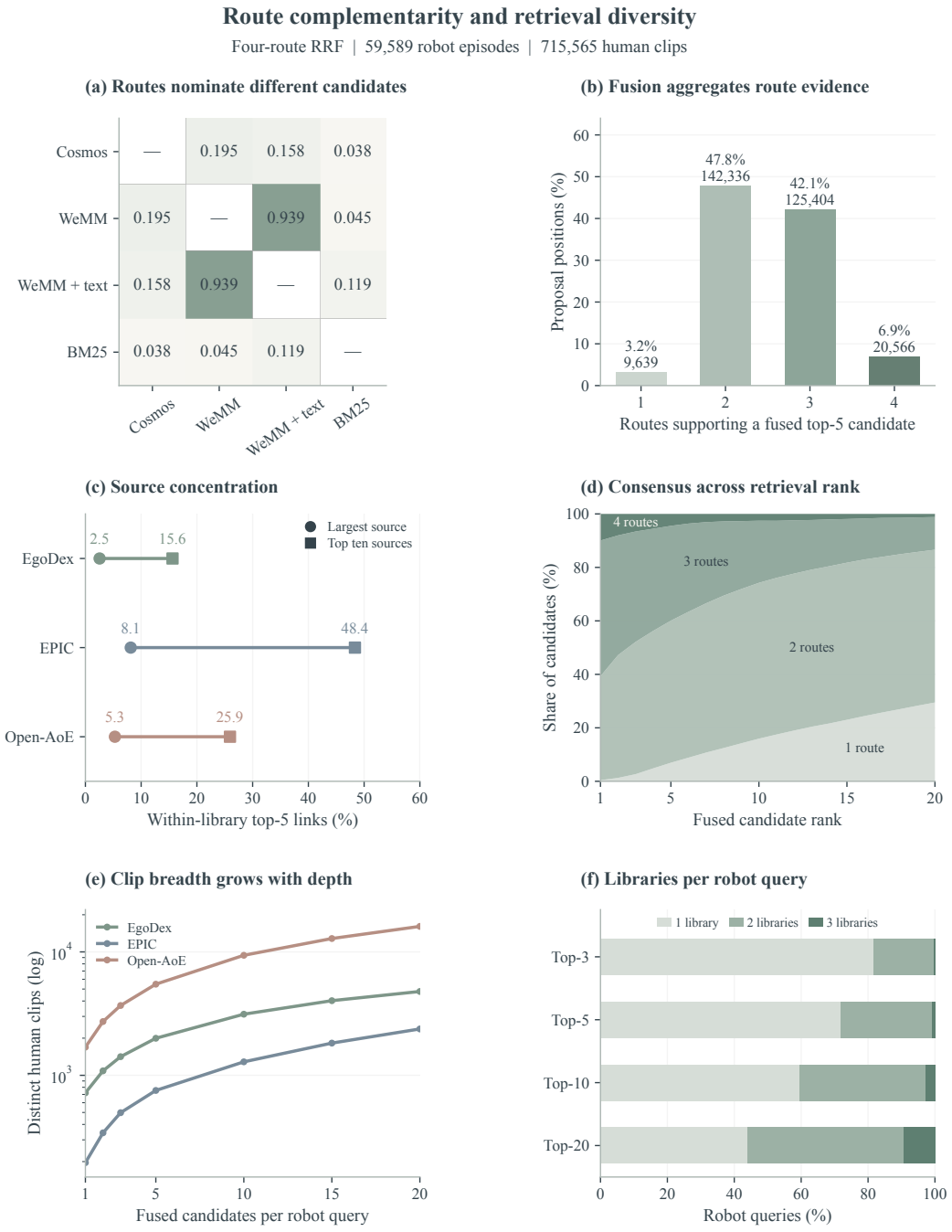



Figure 13: **Complementary nomination routes and human-source diversity.** Full-corpus automatic retrieval. (a) Mean intersection size of two top-5 lists; each diagonal would equal five. (b) Number of routes whose global top-100 contains a fused candidate. (c) Within-library link concentration. (d) Number of supporting routes by fused rank (1–20). (e) Growth in distinct human clips as each robot query retains more candidates. (f) Number of human-video libraries represented per robot query at top 3, 5, 10, and 20.

# D RETRIEVAL EVALUATION AND ALL CONFIGURATIONS

## D.1 FROZEN SCOPES, METRICS, AND ANNOTATION

The four candidate scopes contain 909,829 (Base), 910,213 (Fixed), 909,989 (Action), and 910,373 (Union) windows. Base already contains fixed, action-annotation, and compound windows; the other scopes add the corresponding context proposals. All 84 configurations are evaluated independently within each scope.

Relevance grades are 0 (unrelated/reversed), 1 (related partial correspondence), 2 (functional correspondence), and 3 (strict correspondence). U denotes unresolved relevance. For query $i$ and returned position $j$, functional and strict precision are

$$\mathrm{FP@}\,k = \frac{1}{Nk}\sum_{i=1}^{N}\sum_{j=1}^{k}\mathbf{1}[y_{ij} \geq 2], \qquad \mathrm{SP@}\,k = \frac{1}{Nk}\sum_{i=1}^{N}\sum_{j=1}^{k}\mathbf{1}[y_{ij} = 3]. \tag{13}$$

Coverage counts queries with at least one qualifying position. Unknown and unavailable positions receive no confirmed positive credit at the fixed denominator; unknown optimistic bounds are also computed. Joint functional-and-complete precision additionally requires a separately reviewed complete-process label. In Union, WildWAM's confirmed joint precision is 54.00%, with unknown upper bound 55.33%; event-window-only selection gives 50.67–52.00%.

Relevance and process labels use multiframe inspection under fixed task rules. Returned intervals are reviewed at their exact boundaries; expanded windows do not inherit labels from their seeds. Verifier tiers remain ranking outputs and are not reused as evaluation labels. Shared-pool nDCG uses graded gain $2^y - 1$ and a common reviewed ideal list. The complete table reports functional P@5; the accompanying evidence files preserve strict scores, coverage, unknowns, and process labels.

## D.2 WITHIN-SCOPE BUDGET CONTROLS

For the language/verifier, event-window, explicit-role, and combined conditions, each query retains the same candidate count and 30 verifier slots. WeMM uses eight frames per candidate and eight query frames. Each verifier slot receives its baseline number of sequence frames plus two endpoints; short intervals may repeat sampled frames. Event selection adds metadata comparisons without additional RGB sampling. Base/Fixed/Action use 60,376/60,360/60,397 logical verification frames and 36,823/37,787/37,165 metadata comparisons for the event-selection variants. Cache reuse preserves the logical budget and the frame identities.

For fair comparison on the same dataset and evaluation protocol, we reproduce the original methods as closely as possible, adapting implementations where source code is unavailable and using state-of-the-art estimators on our data to obtain auxiliary inputs required by the original formulations, such as motion or geometric information. STRAP/S-DTW (Memmel et al., 2025) uses fixed-window DINOv2 features (Oquab et al., 2024); R+X-style (Papagiannis et al., 2025), RAM (Kuang et al., 2025), RHyME (Kedia et al., 2025), and WeMM/shared (Zhou et al., 2026b) use restricted candidate pools. SiMDex (Lin et al., 2026) and RoboTok (Qian et al., 2026) proxies match robot-state forward kinematics to sparse human geometry extracted from RGB. Figure 3 uses method names for readability; the implementation distinctions above define the evaluated configurations. Within-system ablations isolate event selection; external rows compare complete evaluated retrieval configurations.

**Representation and alignment baselines.** The full inventory spans global visual similarity, temporal correspondence, and language-conditioned retrieval. Representation rows use CLIP (Radford et al., 2021), LaViLa (Zhao et al., 2023), EgoVideo (Pei et al., 2024), V-JEPA 2.1 (Mur-Labadia et al., 2026), PE AV (Vyas et al., 2026), and Cosmos-Embed1 (NVIDIA, 2025). Temporal and motion-based rows adapt Drop-DTW (Dvornik et al., 2021), HAND (Hong et al., 2026), and importance-weighted retrieval (IWR) (Xie et al., 2025); tracking rows use CoTracker3 (Karaev et al., 2025). TFR is the composed-retrieval baseline from EgoCVR (Hummel et al., 2024), and UniVTG (Lin et al., 2023) supplies the temporal-grounding comparison. These references identify the source methods behind the corresponding local configurations in the complete table.

### D.3 COMPLETE COMPARISON RESULTS

Table 5 groups the retrieval comparisons by method family. The verification variants test how visual evidence changes candidate ordering, while the crop variants test whether focusing on a smaller image region improves retrieval.

**How candidates are ranked.** *Grade ranking* places candidates into discrete relevance categories assigned by the verifier and prioritizes higher categories. *Evidence reranking* instead adjusts the incoming retrieval rank according to supporting or contradicting evidence. These alternatives compare prioritizing a categorical judgment with making evidence-dependent changes to an existing ranking.

**What supports a verification judgment.** *Frame evidence* requires the verifier to identify specific video frames supporting its judgment. *Evidence recheck* adds another check when those frame references are invalid; these references point to video frames, not bibliographic citations. The remaining controls remove the final-state, object-identity, or phase-order checks, or provide additional evidence frames. They isolate which visual evidence contributes to selecting a useful human demonstration.

**How spatial and text controls are compared.** Equal-budget crop variants use the same total RGB-frame budget, redistributing it between full frames and crops; the extra-budget variant adds cropped views. Canonical text rewrites source descriptions into a consistent task formulation before retrieval. These comparisons distinguish the effect of spatial focus or normalized descriptions from changing the ranking strategy.

Table 5: **Retrieval configurations by method family.** Functional P@5 (%). Variants stay together; within each family, rows follow Union score (descending). Verifier scoring, crop budgets, and local baseline implementations are defined in the accompanying text. Colors share the main-figure scale.

| Method / variant | Base | Fixed | Action | Union |
|---|---|---|---|---|
| **WildWAM: event selection and role evidence** | | | | |
| **WildWAM — full** | 61.33 | 60.67 | 62.00 | 62.00 |
| Event windows + verifier | 62.00 | 61.33 | 61.33 | 61.33 |
| Event windows + WeMM | 56.67 | 57.33 | 57.33 | 58.00 |
| Role-aware verifier | 55.33 | 54.67 | 56.67 | 54.67 |
| **Language-augmented candidate retrieval** | | | | |
| BM25 + canonical text; verified | 55.33 | 54.67 | 55.33 | 54.00 |
| BM25; verified | 54.67 | 54.00 | 54.67 | 53.33 |
| Canonical text; verified | 53.33 | 52.67 | 53.33 | 52.00 |
| BM25 + canonical text; WeMM | 48.00 | 48.67 | 48.67 | 48.00 |
| BM25; WeMM | 46.67 | 47.33 | 48.00 | 47.33 |
| Canonical text; WeMM | 46.00 | 46.67 | 47.33 | 46.67 |
| **Multiroute fusion and reranking** | | | | |
| WeMM reranking | 41.33 | 42.00 | 42.67 | 42.00 |
| WeMM reranking; matched budget | 39.33 | 40.00 | 40.00 | 39.33 |
| Reciprocal rank fusion | 26.67 | 26.67 | 28.00 | 26.67 |
| **Verification: scoring and evidence controls** | | | | |
| Base verifier; grade ranking | 52.00 | 51.33 | 52.00 | 50.67 |
| Frame evidence; grade ranking | 43.33 | 44.00 | 45.33 | 44.00 |
| Evidence recheck; grade ranking | 42.67 | 42.67 | 44.00 | 42.67 |
| Evidence recheck; evidence reranking | 42.00 | 42.00 | 43.33 | 42.00 |
| Frame evidence; evidence reranking | 41.33 | 41.33 | 43.33 | 42.00 |
| Evidence reranking; extra evidence frames | 41.33 | 41.33 | 43.33 | 42.00 |
| Evidence reranking; no final-state test | 40.67 | 41.33 | 42.67 | 42.00 |
| Evidence reranking; no identity test | 41.33 | 41.33 | 43.33 | 42.00 |
| Evidence reranking; no phase-order test | 40.67 | 41.33 | 42.67 | 42.00 |

| Method / variant | Base | Fixed | Action | Union |
|---|---|---|---|---|
| Evidence reranking; singleton encoding | 41.33 | 41.33 | 43.33 | 42.00 |
| Base verifier; evidence reranking | 41.33 | 41.33 | 43.33 | 41.33 |
| **Spatial focus and candidate selection** | | | | |
| Full + dynamic crop; extra budget | 42.67 | 42.67 | 42.00 | 42.00 |
| Full-frame control | 41.33 | 42.00 | 42.67 | 42.00 |
| One candidate per proposal group | 41.33 | 42.00 | 42.67 | 42.00 |
| Full + center crop; equal budget | 40.67 | 40.67 | 40.00 | 40.00 |
| Repeated full frames; equal budget | 39.33 | 38.67 | 40.00 | 40.00 |
| Full + dynamic crop; equal budget | 38.67 | 38.67 | 38.67 | 38.67 |
| **Description retrieval and visual fusion** | | | | |
| BM25 descriptions | 45.33 | 45.33 | 45.33 | 45.33 |
| Cosmos + BM25 | 41.33 | 40.00 | 41.33 | 40.00 |
| Cosmos + canonical descriptions | 40.00 | 39.33 | 40.00 | 39.33 |
| Canonical descriptions | 37.33 | 37.33 | 37.33 | 37.33 |
| Cosmos + factor scores | 37.33 | 36.67 | 37.33 | 36.67 |
| Cosmos + original descriptions | 31.33 | 30.67 | 31.33 | 30.67 |
| Action / object / relation factors | 29.33 | 29.33 | 29.33 | 29.33 |
| Original descriptions | 25.33 | 25.33 | 25.33 | 25.33 |
| **Visual and multimodal retrieval baselines** | | | | |
| Cosmos — action-aligned windows | 34.67 | 34.00 | 34.67 | 34.00 |
| Cosmos — full frame | 32.00 | 32.00 | 34.00 | 33.33 |
| Cosmos — re-encoded full frame | 32.00 | 32.00 | 34.00 | 33.33 |
| Cosmos — object crop | 26.67 | 26.67 | 28.00 | 27.33 |
| Cosmos — center crop | 19.33 | 18.67 | 19.33 | 18.00 |
| Cosmos + text — action windows | 22.00 | 21.33 | 22.00 | 21.33 |
| Cosmos + text — compound windows | 22.00 | 21.33 | 22.00 | 21.33 |
| Cosmos + text — fixed windows | 20.00 | 20.00 | 21.33 | 20.67 |
| EgoVideo — full frame | 8.00 | 8.00 | 8.00 | 8.00 |
| EPIC descriptions — text only | 2.67 | 2.67 | 2.67 | 2.67 |
| LaViLa — center crop | 3.33 | 3.33 | 3.33 | 3.33 |
| LaViLa — full frame | 1.33 | 1.33 | 1.33 | 1.33 |
| LaViLa — re-encoded full frame | 1.33 | 1.33 | 1.33 | 1.33 |
| LaViLa — object crop | 0.67 | 0.67 | 0.67 | 0.67 |
| PE AV — video query | 29.33 | 29.33 | 28.67 | 28.67 |
| PE AV — video + text query | 26.67 | 26.67 | 28.00 | 27.33 |
| Qwen — task-focused prompt | 29.33 | 30.67 | 30.00 | 30.67 |
| Qwen — generic relevance prompt | 25.33 | 26.67 | 25.33 | 26.00 |
| V-JEPA 2.1 — full frame | 10.00 | 10.00 | 10.00 | 10.00 |
| WeMM — EPIC text proposals | 40.67 | 42.00 | 42.00 | 41.33 |
| WeMM — video + text query | 40.67 | 41.33 | 42.67 | 41.33 |
| WeMM — video query | 34.67 | 34.67 | 34.00 | 34.00 |
| **Prior retrieval methods: local implementations** | | | | |
| CLIP — text-to-video | 9.33 | 9.33 | 9.33 | 9.33 |
| CLIP — mean visual features | 4.67 | 4.67 | 4.67 | 4.67 |
| DINOv2 — mean visual features | 7.33 | 6.67 | 7.33 | 6.67 |
| Drop-DTW — temporal alignment | 0.67 | 0.67 | 0.67 | 0.67 |
| HAND — detected hand paths | 7.33 | 7.33 | 7.33 | 7.33 |
| HAND — CoTracker3 paths | 5.33 | 5.33 | 6.00 | 6.00 |
| IWR — frozen density score | 0.67 | 0.67 | 0.67 | 0.67 |
| R+X — Qwen relevance | 26.00 | 26.67 | 27.33 | 27.33 |
| RAM — object-region CLIP | 26.00 | 26.00 | 27.33 | 27.33 |
| RHyME — optimal transport | 6.67 | 6.67 | 6.67 | 6.67 |

| Method / variant | Base | Fixed | Action | Union |
|---|---|---|---|---|
| RoboTok — wrist-local geometry | 14.67 | 15.33 | 15.33 | 16.00 |
| RoboTok — body-frame geometry | 10.00 | 10.00 | 11.33 | 11.33 |
| RoboTok — reversed sequence | 8.67 | 8.67 | 10.00 | 10.00 |
| RoboTok — DTW scoring | 8.67 | 8.67 | 9.33 | 9.33 |
| SiMDex — initial pose + text | 12.67 | 12.67 | 13.33 | 14.00 |
| SiMDex — text only | 12.00 | 12.00 | 14.00 | 13.33 |
| SiMDex — add motion + flow | 8.00 | 8.00 | 8.00 | 8.00 |
| SiMDex — add motion | 7.33 | 7.33 | 7.33 | 7.33 |
| SiMDex — motion without fingers | 6.67 | 6.67 | 7.33 | 7.33 |
| STRAP — subsequence DTW | 4.00 | 4.00 | 4.00 | 4.00 |
| TFR — 200-candidate reranking | 10.67 | 10.67 | 10.67 | 10.67 |
| TFR — 15-candidate reranking | 6.00 | 6.00 | 6.00 | 6.00 |
| UniVTG — text confidence | 3.33 | 3.33 | 3.33 | 3.33 |
| WeMM — common candidate pool | 37.33 | 37.33 | 37.33 | 37.33 |

### D.4 Event selection and representation controls

For an event chain, the minimum interval coverage in Equation 2 prevents one well-covered stage from hiding an omitted stage. The harmonic term rewards coverage without an unnecessarily long window. The candidate set preserves the original window and requires same-source event identity. For explicit multi-operation queries, neighboring annotations are considered within 32 seconds, up to the query's required action count; token overlap with the frozen query determines eligible annotated actions.

The four-scope experiment changes the candidate range without changing models, queries, verifier contract, fusion weights, or relevance rules. Event selection improves functional P@5 in every range. Its Union-only result is therefore supported by three additional within-scope comparisons. Explicit roles are most directly reflected in the joint functional-plus-process metric: 54.00% for the combined system against 44.00% for the language/verifier baseline.

Table 6: Representation and computation controls on the fixed top-30 pool, Union scope.

| Representation | Visual tokens | Functional P@5 (%) |
|---|---|---|
| Full frame | 352 | 42.00 |
| Duplicate global, equal budget | 352 | 40.00 |
| Center crop, equal budget | 352 | 40.00 |
| Dynamic crop, equal budget | 352 | 38.67 |
| Dynamic crop, extra budget | 704 | 42.00 |

Dynamic crops use hand-near object detections and a 10% margin. The retained full-frame configuration avoids discarding receiver and interaction context. These controls motivate concentrating additional computation on ordered event evidence rather than spatial cropping. All transformed-feature results use singleton encoding; paired task-cluster intervals and encoding recipes are retained in the evidence package.

## E Reference-learning architecture and experimental details

### E.1 Reference conditioning and supervision scopes

The frozen V-JEPA encoder produces 256 human features of dimension 768, organized into 16 temporal blocks and $4 \times 4$ spatial positions. A trainable encoder maps them to dimension 256, with spatial mixing restricted within each temporal block. Spatial pooling yields 64 direct context tokens. Robot visual, state, and past-action history is processed by a two-layer causal temporal encoder. Its final valid state, normalized and added to scaled learned role queries, reads the visual memory and a null item through four-head attention.

The eight queries allocate pairs to current correspondence, upcoming correspondence, and next-useful correspondence, leaving two additional queries without explicit role supervision. The main soft-gate model averages upcoming/next-useful attention over human time, normalizes away null mass, and applies Equation 5 to the 64 direct visual tokens. All temporal positions retain positive weight. Upcoming role features also supply an action-expert plan token. Video conditioning additionally includes future-goal tokens. Neither future robot observations nor task text enter the T conditioning inputs.

Table 7: Conditioning and supervision in the main reference-learning comparison.

| Component | Internet P | Target-scene T |
|---|---|---|
| Human reference | Retrieved weak pair | Collected strong pair |
| External instruction | Robot task text | Absent |
| Robot context | Current observation + causal history | Same |
| Action/video target | Recorded robot future | Recorded robot future |
| Coarse supervision | Scope, function, compatibility | Not consumed |
| Fine-P variant | Event/coverage/prefix labels | Same T objective as coarse |

Coarse pair labels assign full, partial, none, or unknown correspondence; unknown scores retain a validity mask. Full references enter reference-conditioned control training, while known partial and nonmatching pairs provide scoped relation supervision. Fine labels add human/robot events, local coverage, and causal-prefix mappings. In the coverage-aware variant, the reference-reader next-event endpoint is supervised only when the reference positively covers that event; the robot's own future remains available to its control objective.

### E.2 Data, optimization, and evaluation

The Internet-P training pool contains 270 robot tasks (59,589 episodes) and 100k coarse P links. T uses the strong-paired subset of 202 robot demonstrations with 594 collected human demonstrations and 4,374 admissible strong-reference edges (Appendix A). Internet P never reads the collected T references. The fixed evaluation panel use held-out robot episodes, each with a start and middle window; prediction spans 128 actions and 17 video frames over approximately 4.27 seconds. Auxiliary 4/8/16-second feature targets are distinct from this generated future.

All main configurations initialize from A350000. Trainable parameters comprise 29.91M conditioning parameters, 4.74M auxiliary-head parameters, and 19.66M rank-16 LoRA parameters, totaling 54.31M. The main backbone remains frozen. AdamW (Loshchilov & Hutter, 2019) uses learning rates $10^{-4}$ for conditioning/heads and $10^{-5}$ for LoRA, weight decay 0.01, and gradient clipping at 1. Evaluation uses four denoising steps.

Metrics average windows within each robot, then robots within the panel. Main plots show both training runs rather than treating them as confidence limits. Paired intervals resample robots at fixed trained weights. The two P continuations share nearly all adjacent sampled pairs, so they are repeated training runs rather than independent draws of Internet data. These definitions apply to the reported offline prediction results.

### E.3 Video prediction over task progress

Internet pre-adaptation also improves generated futures under longer target-scene adaptation. This comparison initializes both models from A350000 and holds T at 65,536 updates; WildWAM first receives 65,536 coarse-pair P updates. Evaluation uses the same held-out robot windows and collected human references, with task text absent. Video-curve plots set the first predicted frame to time zero; the 16 future frames span 0–4.0 s on this axis. Figure 14 separates full-frame RGB error from error in the regions that change during the recorded future. The latter uses a shared ground-truth mask: the 20% of pixels with the greatest mean absolute change from the initial frame. It measures appearance prediction in changing regions rather than optical flow.

Over the full prediction horizon, Internet pairs reduce RGB MAE by 2.57% and dynamic-region MAE by 4.20%. During the final segment (3.0–4.0 s), these reductions grow to 3.41% and 4.87%,

respectively. The bread-placement query illustrates a larger benefit at a task-critical transition: at the final predicted frame, RGB error falls by 25.05% and dynamic-region error by 35.15%. The milestone in Figure 7a marks the first sampled frame with the bread transported over the plate (2.93 s relative to the first predicted frame). The query-specific curves show the fixed training run with the same seed; aggregate curves first average windows and both training runs within each robot. Shading denotes paired 95% bootstrap intervals over robot episodes at fixed trained weights.

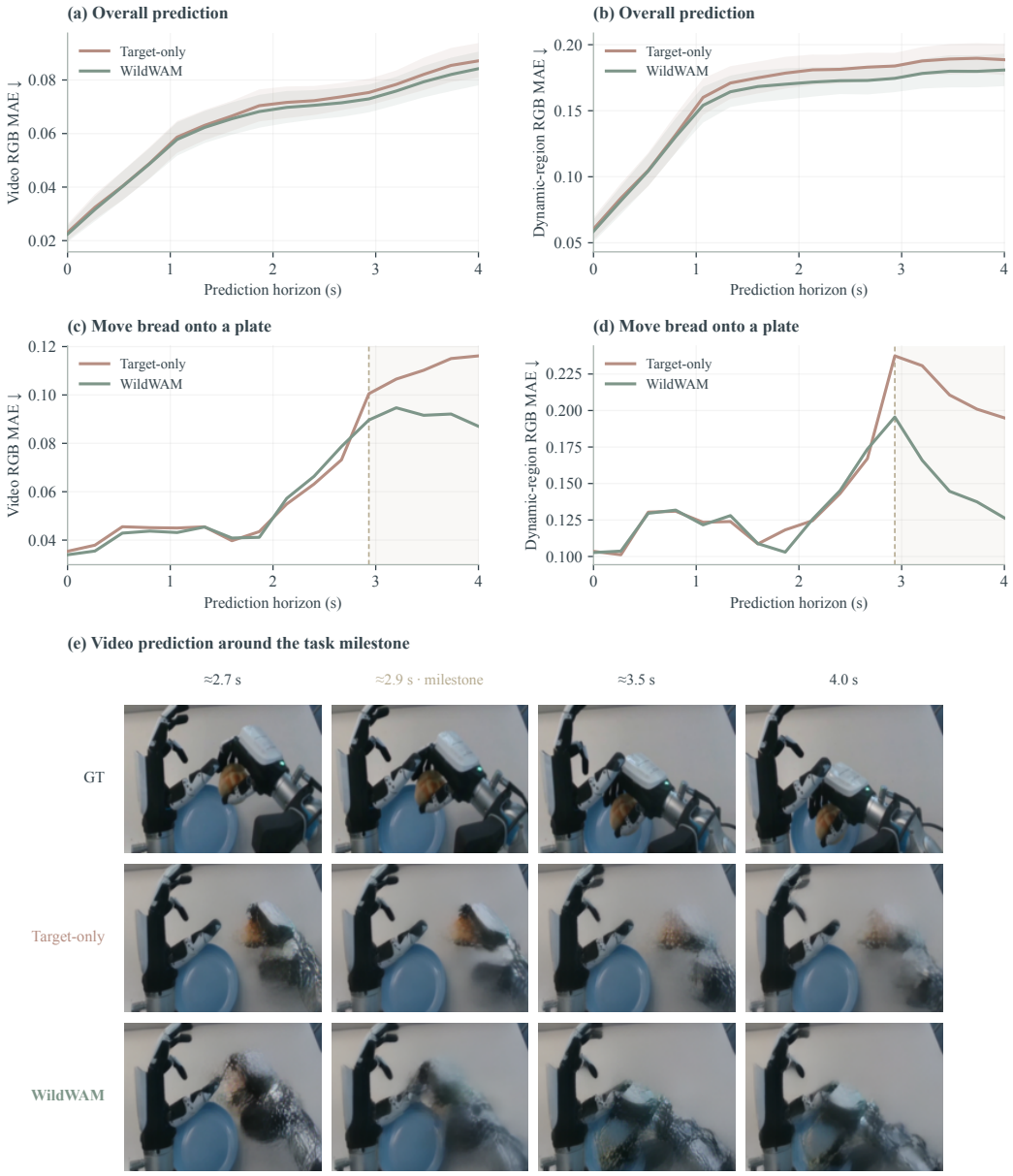



Figure 14: **Internet pairs improve future-video prediction.** Target-only and WildWAM share the target-scene adaptation budget. Full-frame and dynamic-region errors reveal the growing benefit during task progress (here we take the task "move bread onto a plate" as example). In (c,d), the dashed line marks the task milestone and shading denotes subsequent predictions; (e) compares GT, Target-only, and WildWAM around this transition.

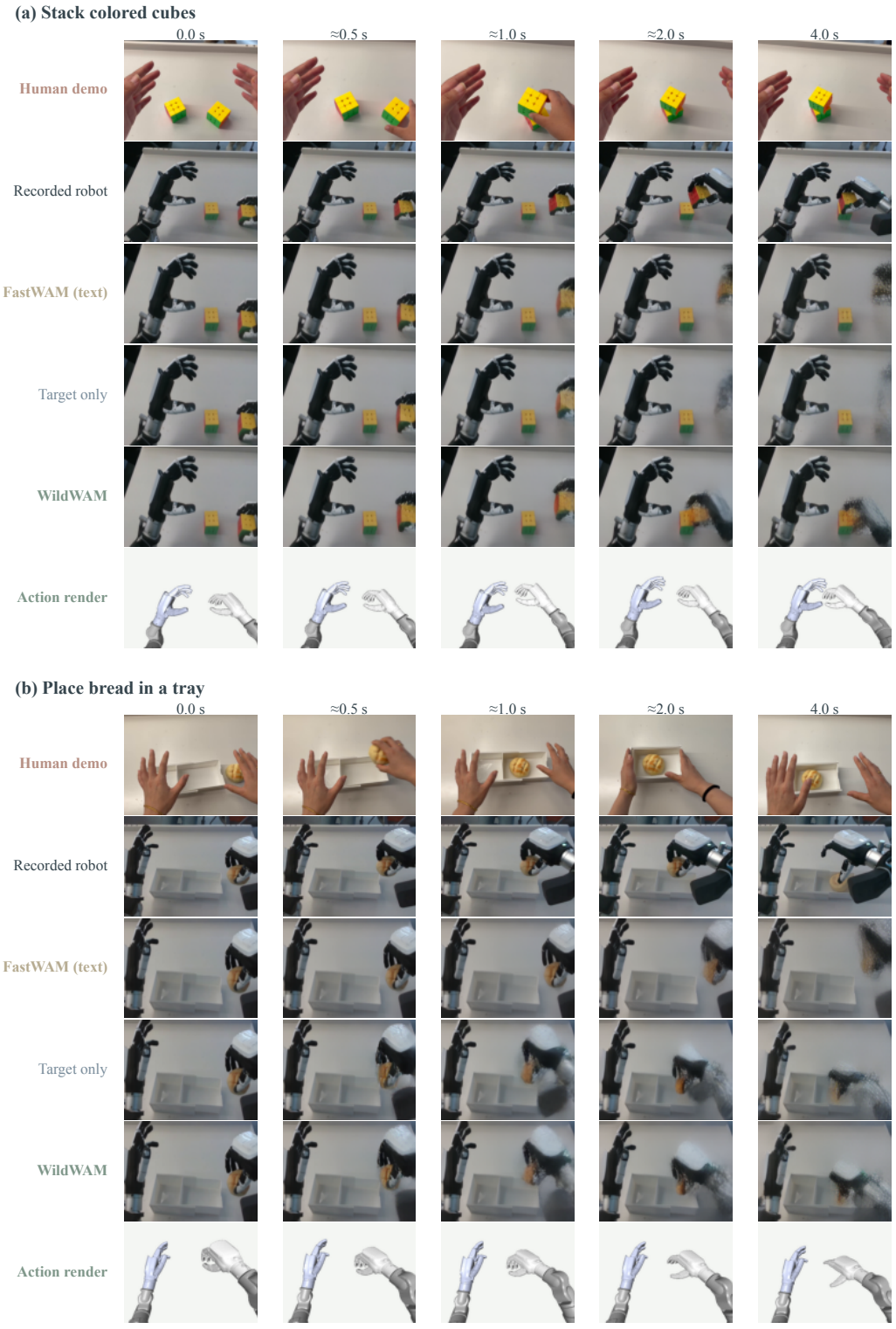



Figure 15: **Additional in-context predictions on different tasks.** Stacking colored cubes and placing bread in a tray. Pretrained FastWAM uses the robot task text and no human reference, which fails to predict futher frames (e.g., the pretrained FastWAM is unable to stack the cube, even though it has predicted the initial picking-up motion.). Target-only adaptation has the similar problem, we attribute this to the lack of Internet human-robot pre-adaptation, which led to overfitting on the small training set. WildWAM predictions use the retrieved Internet human demonstration, enabling a global perspective. Predicted video and G1 + Inspire Hand action renders share the robot horizon.

## F REAL-ROBOT TASK STAGES AND PROGRESS METRIC

**Policy definitions.** Pretrained FastWAM is the robot foundation without P or T adaptation and uses the task text. Internet-only applies P on retrieved Internet human–robot pairs. Target-only applies T on collected target-scene human–robot pairs, starting directly from the foundation. WildWAM applies P followed by T; it therefore transfers Internet procedural evidence before target-scene adaptation.

The five tasks in Table 1 use ordered milestones tied to visible object states and hand–object interactions. The stage count reflects the procedure: a complete placement needs a secure pickup and a stable release, whereas dual placement has two independently completed deposits. Table 1 reports all milestones for each task (15 trials per task and policy).

Table 8: Ordered outcome milestones for the five real-robot tasks. A later stage earns credit only after all earlier stages. "Stable" means that the target relation persists after release; left and right refer to the robot's hands.

| Task | $K_t$ | Milestones |
|---|---|---|
| **Single-arm manipulation** | | |
| Pick-place | 3 | **1** Right hand lifts the object. **2** The object reaches the opening of the container left of center. **3** The object is released and rests stably inside that container. |
| Stack-two | 3 | **1** Right hand lifts the object. **2** It makes supported contact atop the matching object left of center. **3** The released pair remains stably stacked. |
| **Bimanual coordination** | | |
| Dual-place | 2 | **1** The right-hand object is stably deposited in the center container. **2** The left-hand object is subsequently deposited there, with both objects retained. |
| Close-drawer | 3 | **1** Right hand lifts the object. **2** The object is deposited inside the drawer. **3** Left hand braces the drawer while the right hand pushes it fully closed. |
| Replace-press | 4 | **1** Right hand lifts the cleaning tool. **2** The tool is inverted and aligned near the bottle nozzle. **3** Left hand depresses the bottle actuator while the tool remains aligned. **4** Right hand sets the tool down after pressing. |

For trial $i$ of task $t$, let $k_{ti} \in \{0, \dots, K_t\}$ be the longest milestone prefix achieved without human intervention. A reached stage remains credited if a later stage fails. The cumulative stage-success rate and normalized progress are

$$S_{tj} = \frac{100}{N_t} \sum_{i=1}^{N_t} \mathbf{1}[k_{ti} \geq j], \qquad P_t = \frac{100}{N_t K_t} \sum_{i=1}^{N_t} k_{ti} = \frac{1}{K_t} \sum_{j=1}^{K_t} S_{tj}. \tag{14}$$

Here $S_{t,K_t}$ is the final-task success; $P_t$ is the color-encoded stage progress on a common 0–100 scale despite different stage counts. Table 1 displays the aggregate first-stage, second-stage, and final-success rates across all five tasks. For later stages, only tasks defining that milestone enter the denominator (Table 9). The overall final-success and progress scores weight the five tasks equally. The same rubric, initial-scene protocol, time limit, and intervention rule apply to all policies.

Table 9: Aggregate milestone attainment (%). The denominator is 75 trials for stages 1–2 and final completion, 60 for stage 3, and 15 for stage 4. Later-stage rates apply only to tasks with that milestone; progress averages task-normalized stage completion.

| Policy | Stage 1 | Stage 2 | Stage 3 | Stage 4 | Final | Progress |
|---|---|---|---|---|---|---|
| Pretrained FastWAM | 28.0 | 12.0 | 5.0 | 0.0 | 5.3 | 15.0 |
| Internet-only | 10.7 | 1.3 | 0.0 | 0.0 | 0.0 | 4.1 |
| Target-only | 76.0 | 46.7 | 25.0 | 0.0 | 28.0 | 49.4 |
| **WildWAM (Ours)** | **90.7** | **68.0** | **43.3** | **13.3** | **44.0** | **67.7** |

# G AUTOMATIC ANNOTATION FROM COARSE SCOPE TO FINE CORRESPONDENCE

## G.1 CORPUS-SCALE ANNOTATION AND INDEPENDENT OBSERVATIONS

Our corpus-scale automatic annotation covers around **100,000 human videos**, using **GPT-5.6 Sol** (`gpt-5.6-sol`). The schema separates two levels of supervision: coarse descriptions establish what a video demonstrates and how broadly it matches a robot task; fine labels locate the events and state transitions that support transfer. Figure 16 summarizes this process. The laptop–Open-AoE pair below illustrates independent observation, coarse correspondence, and event-level evidence.

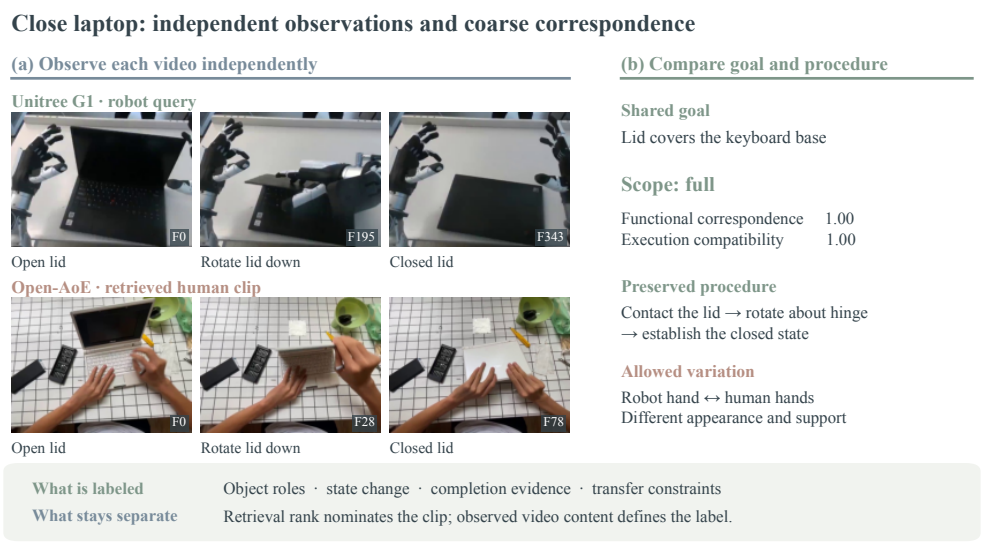



Figure 16: **Coarse annotation of an Internet-retrieved laptop reference.** Robot and Open-AoE videos are observed independently. Both show the lid rotating onto the keyboard base, yielding full correspondence with functional and compatibility scores of 1.00. The labels preserve the shared goal and mechanism while allowing hand and support differences.

Each asset is observed independently before cross-video comparison. Its record includes an action summary, initial and final states, visible entities and roles, and evidence for completion. The observer uses video and temporal coordinates; filenames, retrieval ranks, and the paired embodiment do not supply visual facts. Robot instructions are interpreted separately as requirements, preserving the distinction between a requested goal and an observed outcome.

The fine-annotation protocol samples native frames at 2 Hz, with 8 Hz inspection around uncertain transitions, object identities, and completion states. Long videos are read in 30-second blocks with two-second overlap; events spanning blocks are merged through shared evidence. Sampling retains presentation timestamps and deduplicates repeated frames. Temporal bounds describe the evidence-supported onset and end of an event.

## G.2 COARSE SCOPE AND SCORE SEMANTICS

Coarse pair labels contain scope, functional correspondence, execution compatibility, reusable content, and transfer constraints. Scope takes *full*, *partial*, *none*, or *unknown*. Full requires a complete acceptable procedure and functional/compatibility scores of at least 0.75. Partial identifies specific relevant content without asserting full-program equivalence. Function and compatibility remain separate: one measures the goal relationship; the other measures whether roles, mechanism, and ordering admit transfer.

Table 10: Interpretation of the two coarse correspondence scores. Unknown values retain an explicit mask.

| Score | Functional correspondence | Execution compatibility |
|---|---|---|
| 0 | Unrelated or reversed goal | Mechanism conflict |
| 0.25 | Object or scene association | Clearly unsuitable procedure |
| 0.50 | Related local goal | Substantial procedural adaptation |
| 0.75 | Same function under allowed substitution | Acceptable object, hand, or support changes |
| 1.00 | Object roles and goal agree | Principal roles and order agree |

Asset facts are reused across pair judgments, allowing many robot queries to share a human observation. Pair decisions retain evidence references and reasons; retrieval scores nominate candidates but do not define these labels. Figure 17 shows how a full coarse match can be decomposed into specific event-level support.

### G.3 Events, programs, and local transfer

An event records its source asset, temporal interval and boundary bounds, operation, participants, pre/post states, completion, repeat group, and supporting evidence. Programs retain event IDs, observed ordering, dependency claims, a visible goal, and a terminal state. Outcome criteria express observable relations such as an item resting on a plate, a lid closing, or a drawer moving inward. Local completion is recorded independently of whether another operation begins later in the video.

A correspondence edge may join multiple robot and human events. Its relation is *equivalent*, *transferable with adaptation*, *incompatible*, *appearance-only*, or *unknown*. Transferable edges preserve constraints on objects, tools, geometry, ordering, and coordination. A separate coverage record assigns each robot event *covered*, *absent*, *incompatible*, or *unknown*. This structure preserves useful local evidence within a reference that also contains irrelevant or conflicting material.

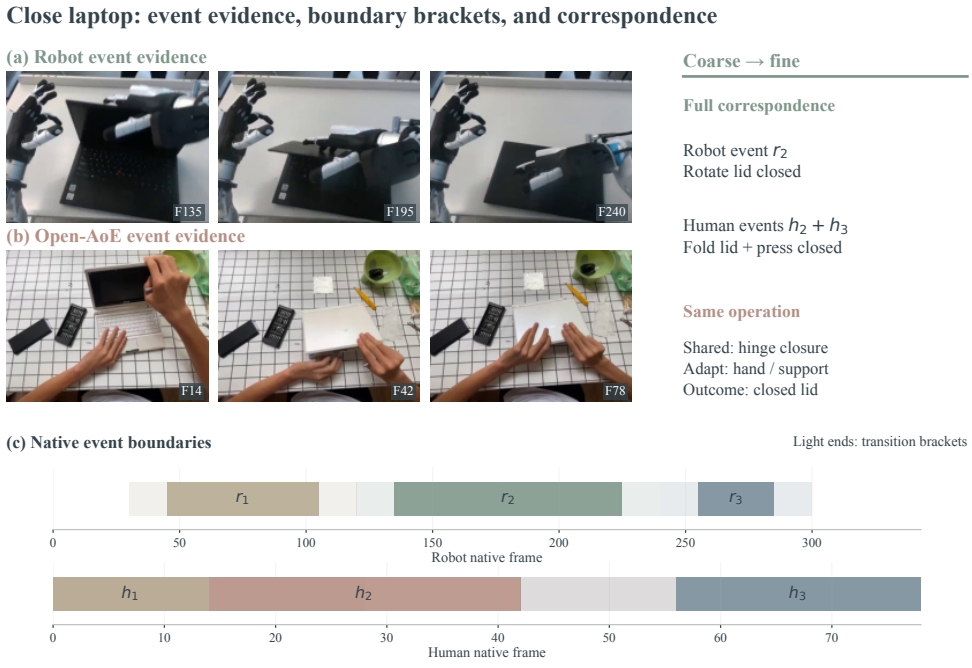



Figure 17: **Fine annotation links closing events across embodiments.** Robot events are contact ($r_1$), rotate closed ($r_2$), and release ($r_3$); human events are grasp ($h_1$), fold the lid ($h_2$), and press closed ($h_3$). The recorded correspondence links $r_2$ to $h_2 + h_3$. Light bar ends show sampled transition brackets. Each timeline retains its own native frame coordinates.

Each semantic claim contains a value, status, evidence IDs, reason codes, and a known mask. Status distinguishes observed, inferred, unknown, and not-applicable fields. Observed/inferred targets can contribute supervision; unknown/not-applicable fields are masked. Event keys include the source asset, preventing repeated local IDs such as `e1` from linking unrelated videos. Processing failures are recorded separately from semantic uncertainty.

### G.4 Task specification, causal prefixes, and future targets

Task-specification records assess whether a human reference and the robot's initial scene define an executable goal. Their states are *sufficient*, *ambiguous*, *infeasible*, *local-skill-only*, and *unknown*. Sufficient references can define a task; local-skill-only references supervise the supported operation. Object-role bindings connect the demonstrated goal to the initial scene.

Causal-prefix annotation reads robot history only through its recorded cutoff. P has instruction-only and instruction-plus-reference conditions; T has no-reference and reference-only conditions without an external instruction. The resulting labels describe visible state, progress hypotheses, and applicable reference events. Future outcomes and prefix-to-event mappings form separate offline targets. The event endpoint and 4/8/16-second horizons each retain availability masks, so a horizon beyond the source video's end contributes no target. Figure 18 reports local relation and coverage proportions for the collected Internet human-robot pairs.

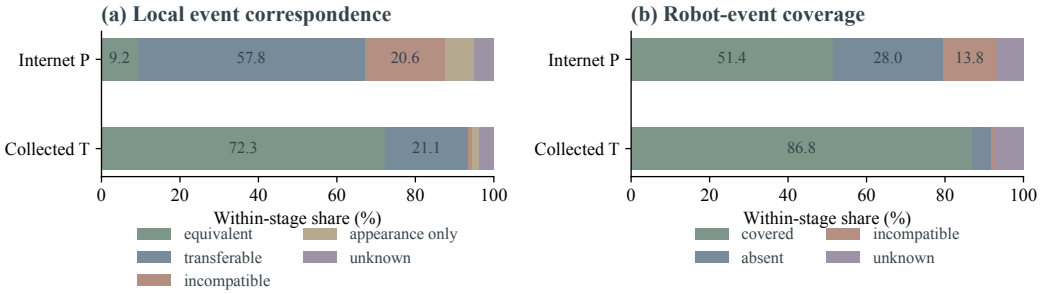



Figure 18: **Fine labels resolve the scope of transfer.** Percentages are normalized within each stage. Local correspondence describes how event pairs relate; coverage describes whether each robot event has supporting human evidence.