From Surveillance to Mobile Robots: Regime-Aware Video Anomaly Detection Supplementary Material

Tatsuya Sasaki*, Yoshiki Ito, Koichiro Yawata, Kota Dohi, Yusuke Ohtsubo, Koki Takeshita
Hitachi, Ltd. Research and Development Group, Japan
{tatsuya.sasaki.gb, yoshiki.ito.xf, koichiro.yawata.rt,
kota.dohi.gr, yusuke.ohtsubo.nb, koki.takeshita.cx}@hitachi.com

This supplementary provides two additional analyses that complement the main paper: **(S1)** a scene/viewpoint-sliced pseudo subdataset analysis examining whether target-level conclusions hold at finer granularity (Tables S1–S2, Figures S1–S3), and **(S2)** an exhaustive subgroup-combination sensitivity analysis quantifying the effect of evaluation subset selection on apparent selector performance (Table S3).

S1. Scene/Viewpoint-Sliced Pseudo Subdataset Analysis

This section supplements the target-level results reported in Tables 3–4 of the main paper. Because target-level averages may hide internal heterogeneity, we run an additional pseudo-subdataset analysis on the same 4-target setting. We split each target into many groups using available scene/viewpoint proxies: `env_id` (Hazards), `physical_scene_id` and `scene_id` (Shanghai), and `sequence_id` (Street/Ped2), plus temporal chunking within each group. The primary-group pool has 52 groups in total (Hazards 3, Shanghai 12, Street 31, Ped2 6).

Protocol. We keep the scoring pipeline unchanged (CLIP-kNN vs. DINOv2-kNN, k=10, PCA $=$ 128) and apply a fixed threshold selector (θ=1.0) from target-train normal frames only. This appendix intentionally uses a fixed threshold (no LOO threshold tuning) to isolate scene-level behavior under a strict no-leak deployment condition. Representative easy/hard normal/anomaly frames in Fig. S3 are selected deterministically from metadata-defined primary groups and are used for qualitative interpretation only (no effect on AUROC computation). All subgroup claims are *within-dataset only*: no cross-dataset subgroup mixing is used. Our default subgroup result uses *all 52 primary groups* (3+12+31+6) with a fixed minimum subgroup size of 100.

Findings.

Hazards is internally consistent: all 3 environments favor DINOv2 and selector gap is zero.

Shanghai is mostly CLIP-favorable but contains minority scenes where DINOv2 is better (worst gap -0.105), indicating residual scene heterogeneity.

Street remains the hardest case: sequence-level winners switch frequently, with the largest miss at -0.284.

Ped2 is generally DINOv2-favorable but still includes sequence-level exceptions (worst gap -0.164).

Figures S1, S2, and S3 visualize the same pattern as distributional evidence.

Implication. Indeed, pseudo subdataset analysis is feasible and informative, and it reveals where regime-aware selection still fails. Accordingly, we treat target-level improvements as an aggregate view and explicitly report scene/sequence-level residual failure modes. For representative subgroup reporting, we use the fixed all-group setting in Table S1 and treat performance-informed subgroup picking as diagnostic-only.

Subgroup selection policy. To avoid ad-hoc subgroup picking, we compare within-dataset selection policies on primary groups (Table S2). For defensive reporting, the representative subset policy is top-n by group size (performance-agnostic). Median-gap and worst-gap are performance-informed policies and are therefore used only as diagnostics in Appendix (not for primary claims).
Chunk granularity. Increasing chunk granularity (2→3→5) consistently worsens mean regret on Street and Ped2, indicating over-fragmentation; we therefore use coarse primary grouping as the default reporting unit and treat fine chunked slices as stress tests only.

S2. Subgroup-Combination Sensitivity Analysis

This section quantifies how evaluation subset selection affects apparent selector performance, complementing the fixed-setting results in Sec. S1. We exhaustively enumerate

Table S1. Pseudo subdataset analysis by scene/viewpoint proxies. Each target is sliced by metadata-derived groups and evaluated under a fixed no-leak threshold (θ=1.0). No subgroup tuple is selected by performance.

Target	Selected Backbone	Global Selector	Valid Groups	Win Rate	Mean Gap (Sel−Best)	Worst Gap
Hazards	DINOv2	0.805	3	1.000	0.000	0.000
Shanghai	CLIP	0.593	12	0.667	-0.016	-0.105
Street	CLIP*	0.439	31	0.548	-0.048	-0.284
Ped2	DINOv2	0.646	6	0.667	-0.039	-0.164
Mean (4 targets)	—	0.621	52 (total)	0.635	-0.037	-0.284

*Fixed threshold (θ=1.0); no override applied.

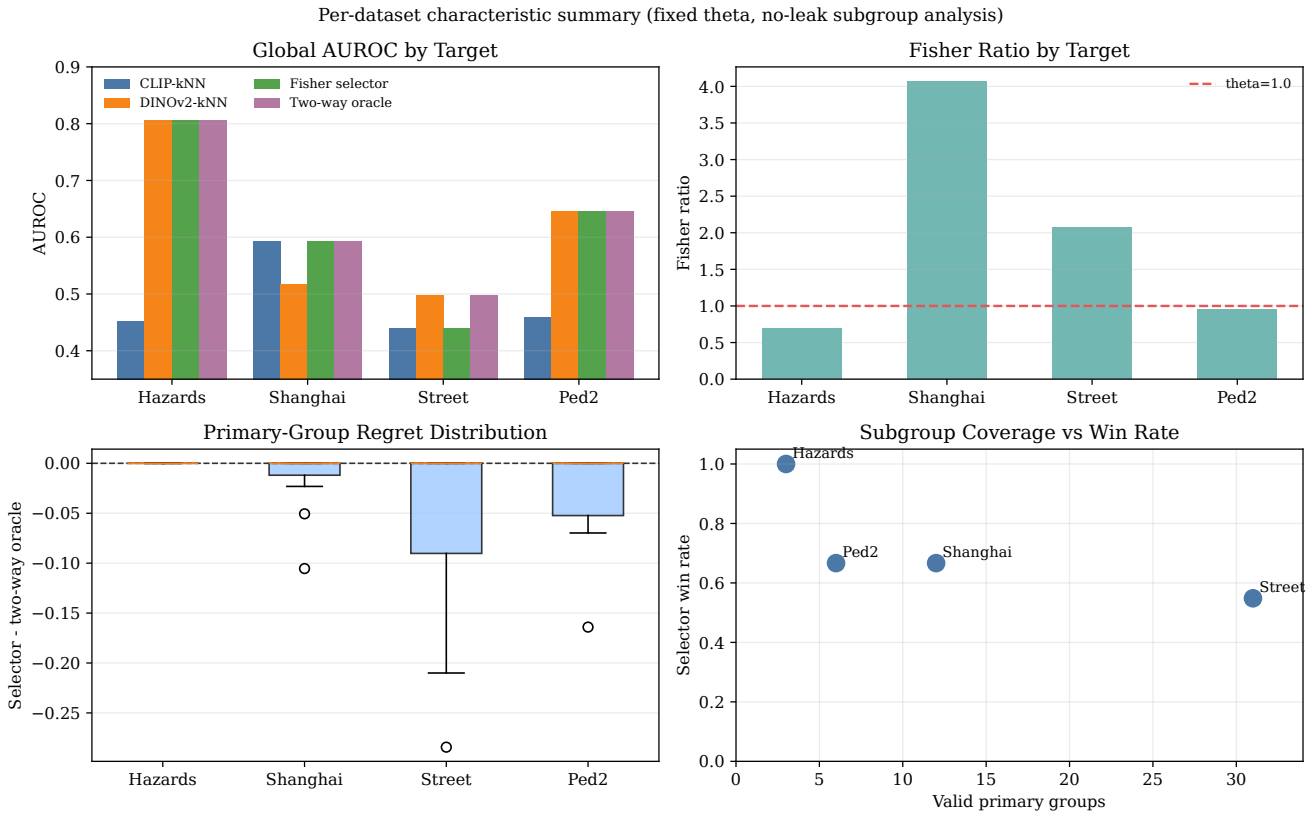

Figure S1. Per-dataset characteristic dashboard under fixed no-leak subgroup analysis ($\theta = 1.0$). Top-left: global AUROC (CLIP-kNN, DINOv2-kNN, Fisher selector, and two-way best). Top-right: Fisher ratio with threshold line. Bottom-left: primary-group regret distribution (selector − best_two_way). Bottom-right: subgroup coverage (valid-group count) versus selector win rate. Hazards is stable (near-zero regret), while Street and Ped2 show heavier negative tails.

all $\binom{3}{3} \times \binom{12}{3} \times \binom{31}{3} \times \binom{6}{3} = 19{,}778{,}000$ subgroup combinations that pick k=3 subgroups per target. We use k=3 because it is the largest uniform choice feasible across all targets (Hazards has exactly 3 primary groups), so the rule is fixed by dataset structure rather than tuned.

Conservative reporting rule. For claims, we use the fixed no-leak all-group result (Table S1, 0.621). Subgroup-combination sweeps are reported only as a sensitivity distribution (Table S3), not as a basis for performance claims.

Results. The distribution center is stable (p50 and mean both 0.695), while the range remains wide (0.595–0.802), indicating that subgroup selection can substantially change apparent performance even when model settings are unchanged. All 19.8M combinations surpass both learned-head baselines; even the worst case (0.595) exceeds the stronger learned head (0.579), confirming the selector benefit is not an artifact of favorable subset choice. Accordingly, we keep strict no-leak target-level metrics as the main claim and treat subgroup-combination results as diagnostic only.

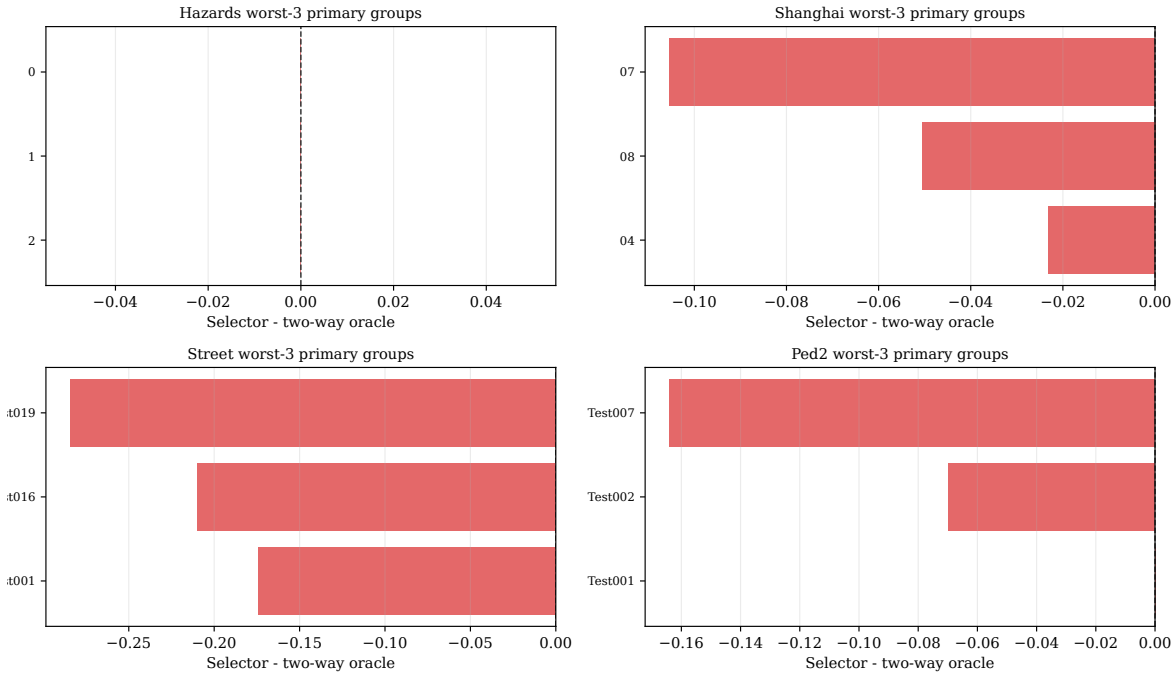

Figure S2. Most difficult primary groups per dataset (lower is worse). Worst-3 primary groups per dataset by selector regret. These slices visualize where residual errors concentrate. Hazards has near-zero worst-group regret, while Street and Ped2 contain larger failure pockets, consistent with Table S1.

Table S2. Subgroup selection policy ablation (k=3, min. size 100). Top-n is performance-agnostic; median-gap and worst-gap are diagnostic-only.

Policy	Mean Selector	Mean Gap (Sel−Best)	Mean Win Rate	Δ vs Fixed ref
Fixed all-group (reference)	0.621	-0.037	0.635	—
top-n (coverage)	0.725	-0.022	0.500	+0.104
median-gap (diagnostic)	0.697	0.000	1.000	+0.076
worst-gap (diagnostic-stress)	0.643	-0.090	0.333	+0.022

Table S3. Subgroup-combination sensitivity (19.8M combinations, k=3 subgroups per target). Δ is measured against the CLIP learned-head baseline (0.579). Best/worst are diagnostic bounds only; see text for details.

Statistic	Selector AUROC	Δ vs CLIP head
Fixed all-52-group result	0.621	+0.042
Combination p10	0.663	+0.084
Combination p50 (median)	0.695	+0.116
Combination mean	0.695	+0.116
Combination p90	0.727	+0.148
Best combination	0.802	+0.223
Worst combination	0.595	+0.016
Head-surpass coverage	100%	—

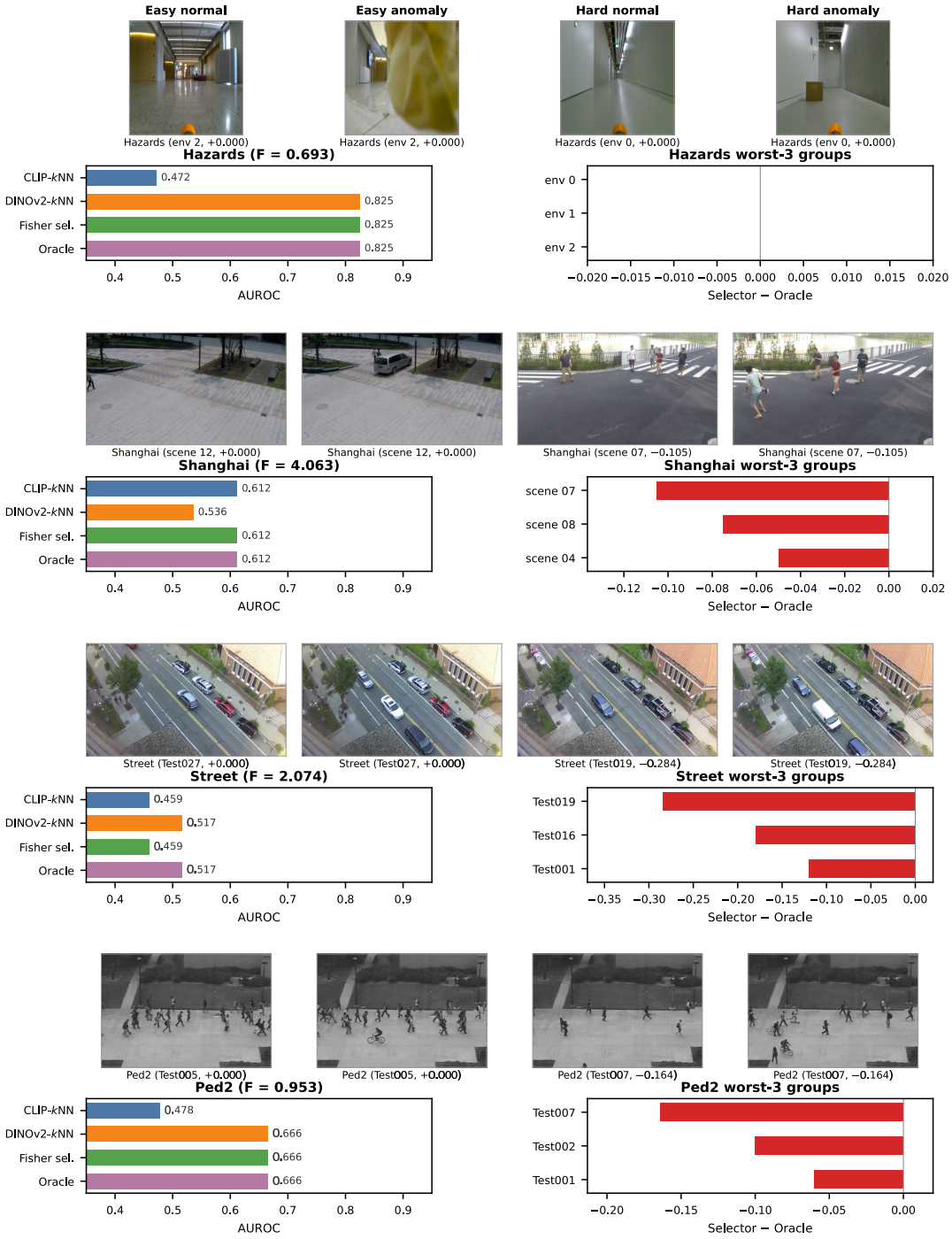

Figure S3. Representative easy/hard subgroup frames with per-dataset quantitative profiles. For each target, we show four representative frames (easy-normal, easy-anomaly, hard-normal, hard-anomaly) under the fixed-threshold selector ($\theta = 1.0$), together with global AUROC bars and worst-3 subgroup regrets. The visual examples are consistent with the distribution-level diagnostics: Hazards is stable, while Shanghai/Street/Ped2 retain hard slices with non-trivial regret.