DINO-Explorer: Active Underwater Discovery via Ego-Motion Compensated Semantic Predictive Coding

Yuhan Jin*, Nayari Marie Lessa*†, Mariela De Lucas Alvarez*, Melvin Laux*†, Lucas Amparo Barbosa‡, Frank Kirchner*†, and Rebecca Adam*
*Robotics Innovation Center, German Research Center for Artificial Intelligence, Bremen, Germany
† Robotics Research Group, University of Bremen, Bremen, Germany
‡ Computing Department, SENAI CIMATEC, Salvador, Brazil
Correspondence: yuhan.jin@dfki.de

***Abstract*—Marine ecosystem degradation necessitates continuous, scientifically selective underwater monitoring. However, most autonomous underwater vehicles (AUVs) operate as passive data loggers, capturing exhaustive video for offline review and frequently missing transient events of high scientific value. Transitioning to active perception requires a causal, online signal that highlights significant phenomena while suppressing maneuver-induced visual changes. We propose DINO-Explorer, a novelty-aware perception framework driven by a continuous semantic surprise signal. Operating within the latent space of a frozen DINOv3 foundation model, it leverages a lightweight, action-conditioned recurrent predictor to anticipate short-horizon semantic evolution. An efference-copy-inspired module utilizes globally pooled optical flow to discount self-induced visual changes without suppressing genuine environmental novelty. We evaluate this signal on the downstream task of asynchronous event triage under variant telemetry constraints. Results demonstrate that DINO-Explorer provides a robust, bandwidth-efficient attention mechanism. At a fixed operating point, the system retains 78.8% of post-discovery human-reviewer consensus events with a 56.8% trigger confirmation rate, effectively surfacing mission-relevant phenomena. Crucially, ego-motion conditioning suppresses 45.5% of false positives relative to an uncompensated surprise signal baseline. In a replay-side Pareto ablation study, DINO-Explorer robustly dominates the validated peak F1 versus telemetry bandwidth frontier, reducing telemetry bandwidth by 48.2% at the selected operating point while maintaining a 62.2% peak F1 score, successfully concentrating data transmission around human-verified novelty events.**

***Index Terms*—Curiosity-Driven Robot Learning, Bio-inspired Robotics, Active Perception, Self-Supervised Learning, Predictive Coding, Autonomous Underwater Vehicles (AUV)**

I. Introduction

Underwater robots are increasingly used for ecosystem observation, scientific survey, and offshore infrastructure inspection, but the resulting visual streams remain costly to review frame by frame [1], [2]. Remotely Operated Vehicles (ROVs) and Autonomous Underwater Vehicles (AUVs) can now collect underwater imagery at scales that are impossible to inspect manually [1]. At the same time, underwater vision remains fundamentally difficult: light scattering, color distortion, turbidity, and marine snow all degrade the sensory stream seen by the robot [2]. These effects are especially severe in the shallow North Sea, where tides, waves, and storms continually resuspend sand and mud from the seabed [3], [4].

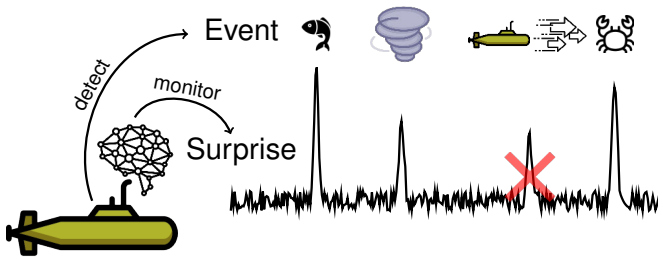


Fig. 1. Conceptual overview of DINO-Explorer: Inspired by predictive coding, the framework generates an intrinsic surprise signal that acts as an indicator of semantic novelty to serve active discovery. Downstream tasks, such as detecting scientifically relevant events (e.g., biological anomalies or habitat transitions), can then be driven by this signal. Furthermore, to mitigate false positives caused by the AUV's own motion, the system isolates genuine environmental changes by implementing an efference copy-inspired ego-motion compensation module.

This creates a perception bottleneck for curiosity-driven underwater robots [5]. A robust novelty indicator for this setting must do more than record long missions: it needs to highlight transient, semantically meaningful changes despite turbidity, illumination drift, marine snow, and frequent camera motion, while remaining stable enough for downstream robot interfaces. Three technical obstacles make this difficult. First, low-level motion estimators based on optical flow [6] or visual odometry [7] are effective for ego-motion estimation, but their geometric and photometric cues are too brittle to define semantic novelty in visually degraded water. Second, a mobile platform continuously generates self-induced visual change, so any novelty signal must separate externally caused events from the sensory consequences of the AUV's own maneuvers. Third, the signal must remain causal and compact enough to be reused for downstream robot decisions rather than optimized only for an offline review dataset.

We address these obstacles by defining *surprise* in semantic latent space. DINO-Explorer draws on two classic ideas from neuroscience and physiology: predictive coding [8], [9], which interprets surprise as deviation from an expected sensory state, and efference copy [10], [11], which explains how self-generated sensory consequences can be discounted. In that sense, DINO-Explorer is a bio-inspired robotics framework: it turns these principles into a practical attention mechanism for an embodied underwater robot. We instantiate these ideas with frozen DINOv3 latent states [12], short-horizon semantic prediction, and ego-motion cues to build a causal surprise signal for underwater robot attention allocation. This signal is not tied to a single downstream task: in robotics it can support selective review, adaptive sensing, and later closed-loop decision modules. This paper evaluates one representative downstream interface: asynchronous underwater event triage under mission-time telemetry constraints.

Specifically, we investigate three core questions: First, can semantic prediction mismatch identify mission-relevant semantic changes despite severe underwater visual noise? Second, can motion conditioning prevent the robot from mistaking its own maneuvers for genuine environmental novelty? Third, when used as a causal trigger policy, can the same surprise signal filter long survey missions and reduce telemetry bandwidth while retaining reviewer-agreed events? We answer these questions by evaluating the downstream trigger policy on offline replay data and a selected operating point against human consensus.

In summary, our primary contributions are threefold:

1) **Motion-aware semantic surprise modeling:** We formulate a lightweight predictive baseline in frozen DINOv3 latent space and use one-step semantic prediction error to define a continuous surprise signal that is more robust than pixel-space novelty cues in noisy underwater video.
2) **Ego-motion-aware false-alarm suppression:** We condition the predictive model on globally pooled optical flow, giving the system an efference-copy-style motion cue that discounts maneuver-induced appearance change instead of treating it as external novelty.
3) **Validation of surprise signal with downstream triage interface:** We treat the continuous surprise score as the core output of the method and evaluate a representative downstream triage interface using a Pareto sweep on replay data, a fixed protocol for human verification, and an analysis of telemetry reduction. The sweep tests the full event-proposal-quality/bandwidth trade-off instead of a single selected threshold; the consensus summary verifies that the selected operating point retains events broadly identified by the human annotators.

II. Related Work

Relevant prior work spans three main areas: active perception and underwater selective discovery, foundation-model representations for semantic perception, and predictive state-space modeling in robotics. Classical active-perception research asks what an embodied agent should sense and when it should spend attention or computation [13], [14]. Most underwater systems, by contrast, still emphasize robust acquisition, mapping, or task-specific inspection rather than online semantic triage [1], [2]. A notable published exception is context-enhanced anomaly modeling for curiosity-driven underwater exploration [5], which targets underwater anomaly-driven discovery more directly than full survey-stack autonomy pipelines.

Foundation-model representations have recently strengthened underwater semantic perception. DINOv2 and DINOv3 features provide a strong semantic substrate for marine downstream tasks [15] [12] . In underwater settings, published studies already show value for expert-assisted marine image annotation protocols [16] and downstream tasks such as instance segmentation with frozen DINO backbones [17]. Together, these results suggest that foundation-model state spaces can remain useful even under severe marine-domain degradation.

Outside the underwater domain, related robotics and world-model literature connects semantic prediction, intrinsic motivation, and self-motion-aware sensing. Curiosity-driven exploration via self-supervised prediction links semantic novelty signals to intrinsic-motivation work [18], while world-model lines from cognitive robotics to latent-dynamics planning and physical robot learning provide a broader conceptual lineage for expectation-based novelty and self-motion-aware sensing [19]–[22]. Recent foundation-model world models such as DINO-WM extend this line to pretrained DINO features by learning dynamics for planning without pixel reconstruction [23]. Practical robotic deployments ground those ideas in deployable motion cues through estimators such as Recurrent All-Pairs Field Transforms (RAFT) and underwater visual odometry [6], [7]. Taken together, these lines motivate continuous surprise not as a task-locked anomaly score, but as a reusable robot-facing signal for attention allocation, selective communication, and later control interfaces. DINO-Explorer evaluates one underwater triage instantiation of that broader robotics role.

III. Preliminary

A. Predictive Coding and Curiosity-Driven Learning

Predictive coding provides the conceptual lens for expectation-based perception: an agent maintains an internal prediction of the next sensory state, and surprise arises when observation departs from that expectation [8], [9]. In curiosity-driven robot learning, the same mismatch can be reused as an intrinsic signal that highlights informative states instead of relying on raw appearance change alone [18], [19]. For robot systems, that kind of surprise signal can gate attention, sampling, or later action selection without committing the representation to a single downstream task.

B. Ego-Motion Compensation and Efference Copy

Efference copy provides the complementary lens for discounting self-generated sensory change [10], [11]. For embod-

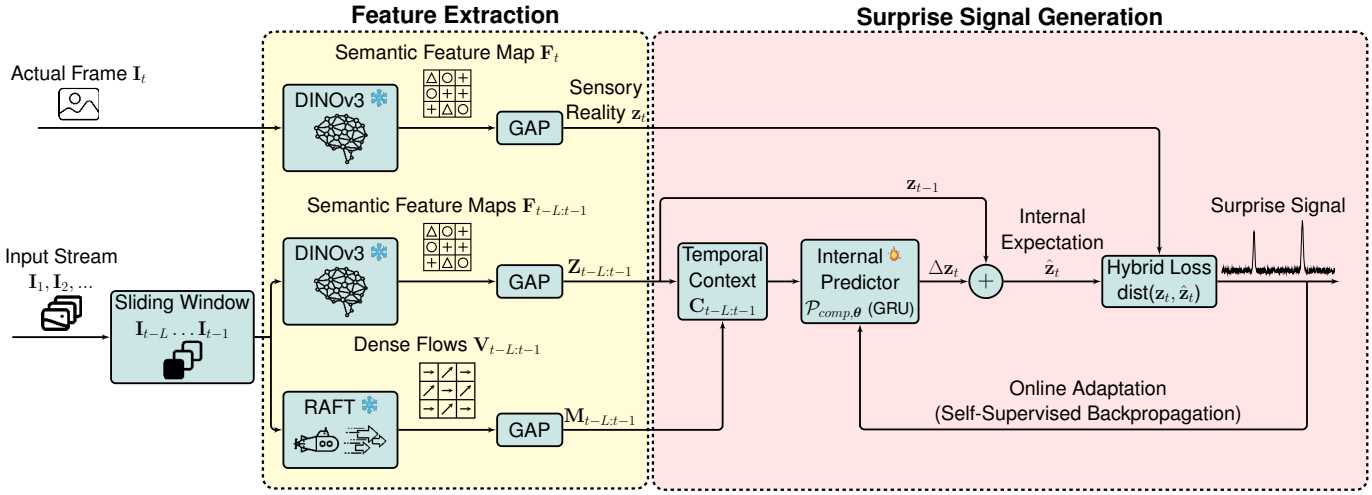


Fig. 2. System architecture of DINO-Explorer: The pipeline comprises *Feature Extraction* (left/yellow), which produces actual semantic state $\mathbf{z}_t$, history semantic states $\mathbf{Z}_{t-L:t-1}$ and RAFT-based motion estimates $\mathbf{M}_{t-L:t-1}$, and *Surprise Signal Generation* (right/pink). Within a predictive coding framework, the GRU-based ego motion compensated semantic predictor ($\mathcal{P}_{comp,\boldsymbol{\theta}}$) generates an internal expectation $\hat{\mathbf{z}}_t$ conditioned on the motion-aware temporal context $\mathbf{C}_{t-L:t-1}$, which acts as an efference copy. The resulting surprise signal $\mathcal{S}_t$, quantifying the mismatch between sensory reality $\mathbf{z}_t$ and internal expectation $\hat{\mathbf{z}}_t$, is used both to drive continuous online adaptation via self-supervised backpropagation and to provide a real-time novelty indicator for downstream active discovery tasks.

ied robots, this principle motivates conditioning semantic expectations on a compact estimate of self-motion, so maneuver-induced transients can be treated as expected reafferent change rather than externally caused novelty. In underwater settings, where direct motor telemetry is often limited or unreliable, optical-flow-based motion cues offer one practical route to that conditioning [6].

Together, these two principles motivate a robot-facing continuous surprise signal: a predictor models expected semantic evolution, and a motion cue discounts robot-induced appearance change.

IV. DINO-Explorer

The Dino-Explorer framework comprises a frozen semantic observation model, a recurrent transition model, an optical-flow-derived action-conditioning module, and a causal event extractor, serving as a downstream task example. The first three components define the internal world model and produce the continuous surprise signal; the last converts that signal into mission-facing review proposals.

Figure 2 presents the Dino-Explorer system architecture, which consists of two primary components. The yellow block represents *Feature Extraction*, which generates the semantic state representation. The red block denotes *Surprise Signal Generation*, where semantic states are processed to produce *Surprise signals* when instances deviate from predicted regularities, as well as the self-supervised backpropagation signal for online module correction.

Unless noted otherwise, in this section, bold lowercase letters denote vectors, bold uppercase letters denote matrices, tensors, or vector sequences, uppercase Roman letters denote fixed dimensions, window lengths, counts, or aggregate energies, and lowercase Roman letters denote indices or per-step scalars. Greek letters denote scalar hyperparameters, and calligraphic letters denote predictors or continuous signals such as $\mathcal{P}$ and $\mathcal{S}$.

A. Semantic State Representation

To strictly focus on global scene semantics, our architecture operates exclusively on a 1D global vector level, as shown in the overall pipeline (Fig. 2). Let $\mathbf{I}_t \in \mathbb{R}^{H\times W\times 3}$ denote the input video frame at time step $t \in \{1, \ldots, T\}$. To overcome the stochasticity of pixel-level underwater noise, we utilize the DINOv3 vision foundation model, denoted as a frozen encoder f_ϕ, to map input frames to a robust semantic latent space. For an input image $\mathbf{I}_t$, the encoder produces a dense feature map $\mathbf{F}_t \in \mathbb{R}^{H'\times W'\times D}$:

$$\mathbf{F}_t = f_\phi(\mathbf{I}_t). \tag{1}$$

where $H' = 32, W' = 32$ are the spatial dimensions of the patch tokens for an input resized to 512×512 pixels, and D is the feature dimension (e.g., $D = 1024$ for ViT-L). To obtain a compact global state representation $\mathbf{z}_t \in \mathbb{R}^D$, we apply global average pooling across the spatial dimensions:

$$\mathbf{z}_t = \frac{1}{H'\times W'} \sum_{i=1}^{H'} \sum_{j=1}^{W'} \mathbf{F}_t^{(i,j)}. \tag{2}$$

This operation collapses the spatial dimensions into a single 1D vector $\mathbf{z}_t$, effectively abstracting away localized artifacts such as caustic lighting or marine snow to focus on global scene characteristics.

B. Predictive Modeling and Naive Surprise Generation

DINO-Explorer uses a Gated Recurrent Unit (GRU) predictor $\mathcal{P}_{\boldsymbol{\theta}}$ as a lightweight internal world model parameterised

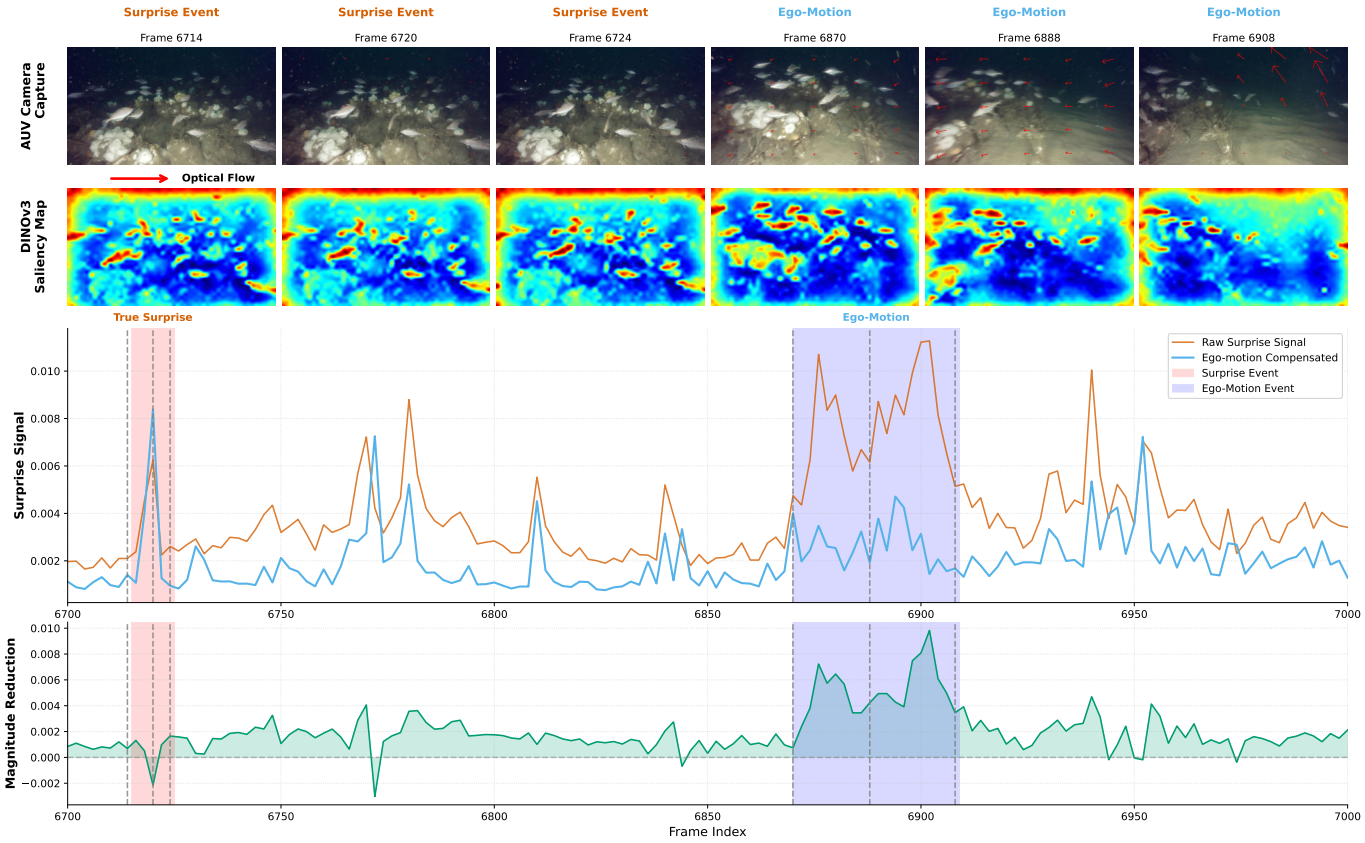


Fig. 3. Qualitative analysis of the semantic surprise signal across a continuous sequence. The plot illustrates the efficacy of our predictive coding framework in isolating genuine novelty: (left/red) DINO-Explorer detects significant deviations from the predicted semantic path caused by a fish rapidly changing its trajectory, generating a valid surprise peak that is correctly preserved after ego-motion compensation; and (right/blue) the efference copy-inspired module effectively filters out predictable reafferent noise induced by an AUV maneuver, suppressing a "naive" false alarm to ensure the autonomous trigger remains robustly focused on scientifically relevant phenomena.

by θ. Given recent semantic states, it predicts the next latent transition one step ahead. The predictor is first pre-trained on continuous underwater footage so it can internalize typical marine dynamics before online adaptation.

The temporal context is represented as a history buffer $\mathbf{Z}_{t-L:t-1} = [\mathbf{z}_{t-L}, \ldots, \mathbf{z}_{t-1}]$ with a lookback window $L = 50$. At each time step t, a GRU processes this buffer via a residual learning formulation to predict the semantic change:

$$\Delta\hat{\mathbf{z}}_t = \mathcal{P}_{\boldsymbol{\theta}}(\mathbf{Z}_{t-L:t-1}), \tag{3}$$

with the predicted semantic state defined as:

$$\hat{\mathbf{z}}_t = \mathbf{z}_{t-1} + \Delta\hat{\mathbf{z}}_t. \tag{4}$$

We quantify the discrepancy between the observed state $\mathbf{z}_t$ and the predicted state $\hat{\mathbf{z}}_t$ using an intrinsic surprise signal as $\mathcal{S}_t$. To capture both magnitude and directional shifts in the latent space, $\mathcal{S}_t$ is calculated with a hybrid loss combining Mean Squared Error (MSE) and Cosine Similarity:

$$\mathcal{S}_t = \|\mathbf{z}_t - \hat{\mathbf{z}}_t\|_2^2 + \lambda(1 - \text{sim}(\mathbf{z}_t, \hat{\mathbf{z}}_t)), \tag{5}$$

where $\lambda = 0.5$ balances the two components.

This surprise signal serves two purposes in this context. First, it serves as an online indicator of environmental novelty, flagging instances in which the scene deviates from predicted regularities (e.g., the sudden appearance of a benthic habitat). Second, it drives continuous self-supervised adaptation. During deployment, the model backpropagates $\mathcal{S}_t$ at every step to refine its weights $\boldsymbol{\theta}$ using a learning rate $\eta = 0.001$:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\nabla_{\boldsymbol{\theta}}\mathcal{S}_t(\mathbf{z}_t, \hat{\mathbf{z}}_t(\boldsymbol{\theta}_t)). \tag{6}$$

This iterative update enables the internal world model to track gradual contextual drift while maintaining high sensitivity to transient anomalies. Therefore, this mechanism effectively isolates novelty, such as the sudden trajectory change of a fish, by identifying significant deviations from the predicted semantic path.

C. Ego Motion Compensation with Optical Flow

To mitigate false positives induced by vehicle maneuvers, we incorporate an action conditioning module grounded in *efference copy*. Since precise motor telemetry is often unavailable or drift-prone in underwater settings, we approximate the agent's motor state using dense optical flow $\mathbf{V}_t$ computed via a RAFT estimator [6]. We derive a scale-invariant 2D global

translation vector $\mathbf{m}_t = [\bar{v}_x/W, \bar{v}_y/H] \in \mathbb{R}^2$ by spatially averaging the x and y components of the flow field (denoted as $\bar{v}_x$ and $\bar{v}_y$) and normalizing the result by the image width W and height H.

This global pooling mechanism ensures that the motion cue primarily captures the overall background shift caused by the vehicle rather than the localized movement of individual objects. While moving fauna generate isolated flow vectors, the background, which dominates the frame, exhibits a consistent global shift corresponding to the AUV's trajectory. Averaging the flow across the frame yields a robust proxy for ego-motion with reduced sensitivity to transient subject dynamics.

We formulate the ego motion compensated predictor $\mathcal{P}_{comp,\boldsymbol{\theta}}$ as a conditional GRU. At each time step t within the lookback window L, the global semantic state $\mathbf{z}_t \in \mathbb{R}^D$ is concatenated with the 2D global motion vector $\mathbf{m}_t \in \mathbb{R}^2$ to form $\mathbf{c}_t = [\mathbf{z}_t, \mathbf{m}_t] \in \mathbb{R}^{D+2}$. The temporal sequence $\mathbf{C}_{t-L:t-1} = [\mathbf{c}_{t-L}, \ldots, \mathbf{c}_{t-1}]$ is processed by a two-layer GRU with hidden dimension 256, allowing the predictor to model how vehicle motion shifts the high-level semantic state.

The final hidden state is mapped back into the semantic space by a fully connected linear layer, yielding the motion-compensated predicted semantic change:

$$\Delta\hat{\mathbf{z}}_{comp,t} = \mathcal{P}_{comp,\boldsymbol{\theta}}(\mathbf{C}_{t-L:t-1}). \tag{7}$$

The compensated internal expectation $\hat{\mathbf{z}}_{comp,t}$ is subsequently formed by applying this anticipated shift to the prior semantic state $\mathbf{z}_{t-1}$:

$$\hat{\mathbf{z}}_{comp,t} = \mathbf{z}_{t-1} + \Delta\hat{\mathbf{z}}_{comp,t}. \tag{8}$$

Following this motion-conditioned prediction, the framework proceeds identically to the uncompensated naive surprise signal. The compensated surprise score $\mathcal{S}_{comp,t}$ is calculated using the exact same hybrid loss (Eq. 5) between $\mathbf{z}_t$ and $\hat{\mathbf{z}}_{comp,t}$, and is similarly backpropagated at every step to drive continuous online adaptation. By effectively filtering out predictable reafferent noise, this compensation module mitigates false alarms induced by AUV maneuvers, ensuring the autonomous trigger remains robustly focused on pure semantic dynamics (Fig. 3, right/blue).

D. Downstream Event Extraction for Review and Telemetry

The core output of DINO-Explorer is the ego motion compensated continuous surprise signal $\mathcal{S}_{comp,t}$. One downstream use of this signal, and the one evaluated in this paper, is to convert it into discrete event proposals for human review and telemetry escalation. We instantiate the causal proposal extraction interface in three steps.

First, we smooth the online signal with a one-sided Gaussian kernel using only current and past samples. With truncation radius $K = \max(1, \lfloor 3\sigma \rceil)$ and weights $w_k = \exp(-k^2/2\sigma^2)$, the causal smoother is

$$\bar{\mathcal{S}}_t = \frac{\sum_{k=0}^{\min(t,K)} w_k \mathcal{S}_{t-k}}{\sum_{k=0}^{\min(t,K)} w_k}. \tag{9}$$

Second, because underwater background variance is strongly non-stationary, we normalize against a trailing context window

$$\Omega_t = \{i : \max(T_{\text{warmup}} + 1, t - W_{\text{samp}}) \le i \le t\}, \tag{10}$$

and define the local statistics and adaptive threshold as

$$\begin{aligned} \mu_t &= \text{mean}_{i\in\Omega_t}\, \bar{\mathcal{S}}_i, \qquad s_t = \text{std}_{i\in\Omega_t}\, \bar{\mathcal{S}}_i, \\ \tau_t &= \max(\mu_t + \alpha s_t, \tau_{\min}). \end{aligned} \tag{11}$$

where W_{samp} is the sample count corresponding to the time horizon W_{sec}.

Third, a review proposal is emitted only at causal local maxima that also exceed the adaptive threshold:

$$t \in \mathcal{T} \iff t > T_{\text{warmup}} \wedge \bar{\mathcal{S}}_{t-1} < \bar{\mathcal{S}}_t \ge \bar{\mathcal{S}}_{t+1} \wedge \bar{\mathcal{S}}_t > \tau_t. \tag{12}$$

To prevent duplicate alarms for the same macroscopic phenomenon, accepted peaks are greedily filtered with a refractory interval R. Together, $(\sigma, W_{\text{sec}}, \alpha, \tau_{\min}, T_{\text{warmup}}, R)$ define a family of causal extraction policies built on the same surprise signal. In Section V-D, we specify the operational point used for human review and all reported metrics.

V. Evaluation Setup

The evaluation follows the three research questions posed in the introduction: whether trigger proposals surface review-relevant events, whether motion conditioning suppresses maneuver-induced false alarms, and whether the trigger stream effectively reduces telemetry burden while perceive semantic novelty. We design the evaluation in two levels. The fixed operating point asks what the causal event extractor sends to reviewers under the threshold used for the annotation workflow. The later α sweep pareto front analysis asks a broader question: how the same extractor behaves as its sensitivity varies. This avoids judging the method only at a single selected threshold and provides an overview benchmark across the event-quality and telemetry-bandwidth trade-off.

A. Dataset

We evaluate DINO-Explorer on 178.2 minutes of 720p underwater footage collected during native-oyster restoration surveys at Borkum Riffgrund in the North Sea [24]. The data were recorded with a BlueROV2 [25] at speeds up to 1.6 m/s and depths up to roughly 40 m, and include sandy and coarse substrates, shell accumulations, reef structures, benthic fauna, and frequent suspended sediment.

B. Surprise Event Taxonomy

To operationalize the current human-review target, we group relevant events into three categories:

- *Spatial Transitions*: scene switches, frame entry/exit, and object discoveries.
- *Environmental Events*: lighting changes and turbidity bursts that alter visibility.
- *Animal & AUV Behavior*: meaningful behavioral shifts rather than simple appearance changes.

This taxonomy defines the operational annotation target for the study. We focus on events large enough to disrupt overall scene semantics and ignore steady-state background dynamics and minor local motion. Representative examples, except the Animal Behavior which is already shown in Fig. 3, are demonstrated in Appendix A.

C. Human Annotation and Verification Protocol

Defining novelty in continuous video is inherently subjective and prone to cognitive fatigue in humans (e.g., inattentional blindness during long sequences without events). To account for this bias and establish a fair human consensus ground truth, we conducted a two-phase verification study involving 15 human reviewers spanning different ages and backgrounds:

- **Phase 1: Blind annotation.** Reviewers mark interval-level novelty events without access to model outputs.
- **Phase 2: Model-guided validation.** Reviewers assess clips centered on automated triggers t^* that fall outside their Phase 1 intervals and mark each as *Agree* or *Reject*.

This yields two annotation-derived event sets: Phase 1 consensus events from blind annotation and model-guided confirmations from reviewer assessment of the surprise signal proposals.

D. Operational Trigger Configuration

All human reviewed clips and operating-point tables use the downstream extractor from Section IV-D. Unless noted otherwise, both model variants use one-sided Gaussian smoothing with $\sigma = 2.0$, a trailing threshold window of $W_{\text{sec}} = 10.0$ s, sensitivity $\alpha = 2.5$, minimum threshold $\tau_{\min} = 0.005$, warmup horizon $T_{\text{warmup}} = 200$ frames, and refractory interval $R = 0.5$ s. These parameters are used during annotation and define a stable scene-level operating point for human review; the replay Pareto analysis then varies α to characterize the surrounding extraction family.

E. Evaluation Metrics

The metrics follow the two-level evaluation design above. The fixed operating point reports how the extracted surprise-event proposals align with reviewer consensus at the selected extractor threshold. The α sweep reports how the same extractor behaves across sensitivity settings, so the comparison is not tied to a single selected operating point.

a) Fixed operating point: At the fixed operating point, Phase 1 Subject-Pool Consensus Recall (SPCR) is the recall-side human-reviewer consensus-retention metric: it measures how many Phase 1 category-consensus events are retained by the operating-point trigger stream. We also report a post-discovery SPCR for the fixed operating point. This companion value adds model-guided Phase 2 discovery confirmations. Phase 1 Trigger Confirmation Rate (TCR) is the precision-side proposal-confirmation metric: it measures the fraction of extracted proposals validated by the Phase 1 consensus. Post-discovery TCR keeps the denominator fixed as all extracted proposals at the operating point, but expands the numerator to include proposals confirmed during model-guided Phase 2 review. Discovery Rate (DR) measures additional annotator-validated events surfaced beyond Phase 1 blind review. False Positive Suppression Rate (FPSR) isolates how much ego-motion compensation reduces false alarms relative to the uncompensated predictive baseline:

$$\mathrm{FPSR} = 1 - \frac{N_{comp}^{false}}{N_{uncomp}^{false}}, \tag{13}$$

where N_{comp}^{false} and N_{uncomp}^{false} are the corresponding false-alarm counts under the operating-point extraction rule.

b) α-sweep replay benchmark: Across the α sweep, validated peak F1 is the main event-proposal-quality measure. We match extracted trigger peaks one-to-one to reviewer-validated proposal peaks within the tolerance window and summarize the resulting peak precision and peak recall with

$$\mathrm{F1}_{\text{peak}} = \frac{2\,P_{\text{peak}}\,R_{\text{peak}}}{P_{\text{peak}} + R_{\text{peak}}}, \tag{14}$$

where P_{peak} and R_{peak} denote matched-peak precision and recall. Phase 1 SPCR and Phase 1 TCR are reported in the sweep as diagnostic recall- and precision-side views against the blind Phase 1 consensus.

c) Telemetry bandwidth: Bandwidth Savings Ratio (BSR) measures telemetry reduction relative to continuous 30 frames-per-second (FPS) streaming. The BSR computation assumes 1 FPS default streaming and 30 FPS inside a ± 3 s window around each trigger:

$$\mathrm{BSR} = 1 - \frac{N_{\text{tx}}}{N_{\text{raw}}}, \tag{15}$$

where N_{raw} denotes the total raw-frame count and N_{tx} the number of transmitted frames under this transmission rule.

d) Latent-energy retention: Latent Energy Retention (LER) provides a replay-side diagnostic for the surprise-trigger stream that does not depend on human event labels. We first compute a reference semantic-change trace from the full video in the shared frozen DINOv3 latent space. Each trigger policy is then scored by the fraction of this reference trace retained by its selected telemetry windows. This keeps the annotation process and representation backbone fixed, so differences in LER reflect the trigger policy's ability to preserve DINOv3-semantic changes under downsampling. Given the frozen DINO latent sequence $\{\mathbf{z}_t\}_{t=1}^{T}$, let

$$\boldsymbol{\mu}_t = \frac{1}{C}\sum_{k=t-C}^{t-1}\mathbf{z}_k, \qquad \delta_t = 1 - \frac{\mathbf{z}_t^{\top}\boldsymbol{\mu}_t}{\|\mathbf{z}_t\|_2\,\|\boldsymbol{\mu}_t\|_2}, \tag{16}$$

where C is the rolling context length and $\delta_t = 0$ when either norm is zero. We then median-center the cosine deviation and clip it to the non-negative part,

$$e_t = \max\Big(\delta_t - \mathrm{median}\{\delta_j\}_{j=C}^{T},\, 0\Big), \tag{17}$$

which defines the per-frame latent-change energy. Given the trigger-induced telemetry mask w_t, summing over the replay gives

$$E_{\text{total}} = \sum_{t=1}^{T} e_t, \qquad E_{\text{captured}} = \sum_{t=1}^{T} w_t e_t, \tag{18}$$

where $w_t = 1$ inside trigger windows and $w_t = r_{\text{low}}$ elsewhere, with r_{low} denoting the low-rate sampling ratio. We report

$$\text{LER} = \frac{E_{\text{captured}}}{E_{\text{total}}} \times 100, \tag{19}$$

so a higher LER indicates greater retention of the reference latent-change trace in the downsampled telemetry stream.

VI. Results

TABLE I
OPERATING-POINT CONSENSUS SUMMARY. PHASE 1 (P1) USES BLIND SAME-CATEGORY CONSENSUS INTERVALS; P1+P2 ADDITIONALLY COUNTS REVIEWER-CONFIRMED MODEL-GUIDED DISCOVERIES. FPSR COMPARES AGAINST THE UNCOMPENSATED PREDICTIVE VARIANT, AND DR REPORTS ADDITIONAL EVENT COVERAGE WITH MODEL DISCOVERING.

Setting	SPCR (%)		TCR (%)		FPSR (%)	DR (%)
	P1	P1+P2	P1	P1+P2		
$\alpha = 2.5$	47.5	78.8	11.8	56.8	45.5	147.8

A. Reviewer-consensus validation at the operating point

Table I summarizes reviewer consensus at the fixed $\alpha = 2.5$ operating point used for the annotation workflow. The 47.5% Phase 1 SPCR means that the trigger stream retained 47.5% of the blind same-category consensus events. After adding reviewer-confirmed Phase 2 model-guided discoveries, post-discovery SPCR rises to 78.8%, indicating broader event coverage once the model-guided review stage is included. The 11.8% Phase 1 TCR means that 11.8% of extracted proposals were confirmed by the blind Phase 1 consensus, while the 56.8% post-discovery TCR means that 56.8% of proposals were confirmed after including Phase 2 reviewer decisions. The 45.5% FPSR means that ego-motion compensation reduced false alarms by 45.5% relative to the uncompensated predictive variant under the same extraction setting. The 147.8% DR means that model-guided review surfaced additional events missed during Phase 1 blind annotation; these events were proposed by the model at Phase 2 and then confirmed by majority reviewer vote in Phase 2.

B. Replay pareto ablation benchmark

Fig. 4 evaluates the extractor over an α sweep, rather than only at the fixed human-review setting in Table I. Each point corresponds to a trigger stream produced at one sensitivity setting, and the curves show how proposal quality and diagnostic retention change as telemetry bandwidth varies. The comparison includes the compensated model, the uncompensated predictive variant, Direct $\Delta \mathbf{z}_t$ as a non-predictive frame-to-frame DINO-latent-change baseline, and Uniform periodic sampling as a content-agnostic baseline that selects telemetry windows at fixed time intervals.

Panel (a) shows that DINO-Explorer forms the strongest validated-peak-F1 frontier, with Direct-Diff as the closest simple baseline, Naive Surprise below it, and Uniform periodic sampling as the weakest reference. This indicates that DINO-Explorer improves the quality-bandwidth trade-off across extraction sensitivities: for a comparable telemetry budget, it yields trigger proposals that better align with reviewer-confirmed events than raw latent change, the uncompensated surprise signal, or plain periodic sampling.

Panel (b) provides the Phase 1 SPCR recall-side view against blind consensus and shows that DINO-Explorer stays near or above the recall-friendly Direct-Diff and Naive Surprise baselines, indicating that ego-motion compensation filters false alarms while retaining the core blind-consensus events annotated by reviewers.

Panel (c) is mainly a diagnostic precision-side view: it shows how often the triggered proposals are confirmed by the Phase 1 blind-consensus reference. The similar Phase 1 TCR curves indicate that this panel is auxiliary to the main peak-F1 frontier in Panel (a), rather than the primary source of DINO-Explorer's contribution.

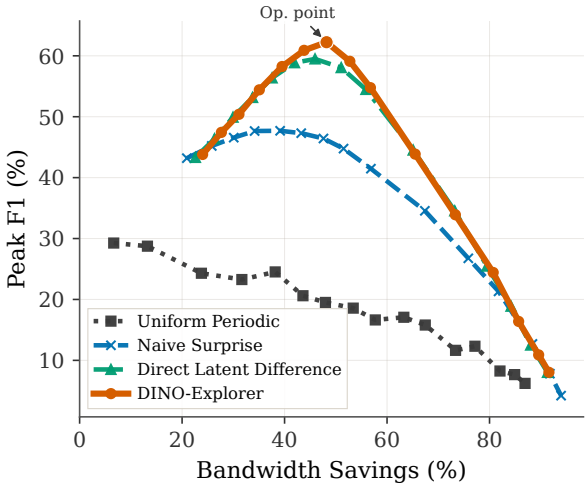
Panel (d) shows that DINO-Explorer remains near or above Direct-Diff in LER, even though Direct-Diff is a strong latent-energy baseline because it directly tracks frame-to-frame change in the same DINOv3 latent space. DINO-Explorer also stays slightly above Naive Surprise over much of the range. This suggests that DINO-Explorer filters motion-induced residuals without losing the main DINOv3-semantic dynamics. When the compensated model exceeds the naive variants, it indicates that motion-conditioned prediction may help concentrating the retained telemetry around semantically informative dynamics beyond raw frame-to-frame latent differences.

TABLE II
OPERATING-POINT REPLAY COMPARISON. PEAK F1 IS MEASURED AGAINST REVIEWER-CONFIRMED PROPOSAL PEAKS; BSR, SPCR/TCR, AND LER REPORT BANDWIDTH, RECALL/PRECISION DIAGNOSTICS, AND LATENT-ENERGY RETENTION. BOLD VALUES MARK THE PRIMARY DINO-EXPLORER OPERATING-POINT READOUTS.

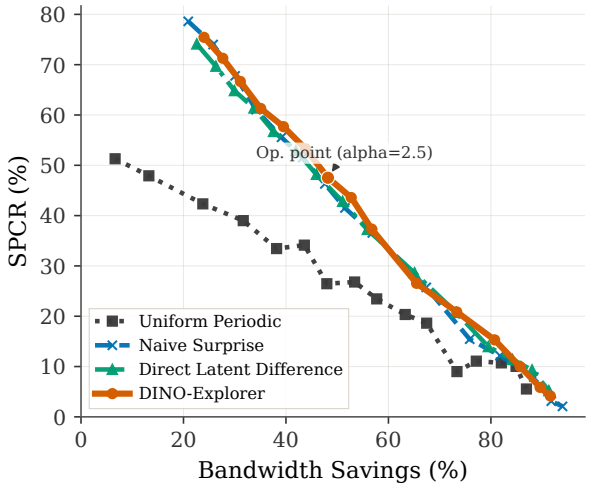
Method	Setting	BSR (%)	Peak F1 (%)	SPCR P1 (%)	TCR P1 (%)	LER (%)
$\mathcal{S}_{comp}$	$\alpha = 2.5$	**48.2**	**62.2**	**47.5**	11.8	**70.8**
$\mathcal{S}_{uncomp}$	$\alpha = 2.5$	47.6	46.4	46.3	12.0	68.2
Direct diff ($\Delta \mathbf{z}_t$)	$\alpha = 2.5$	45.9	59.5	48.2	11.3	71.2
Uniform periodic	$\Delta t = 12$ s	48.0	19.5	26.5	8.6	53.0

C. Replay benchmark at the operating point

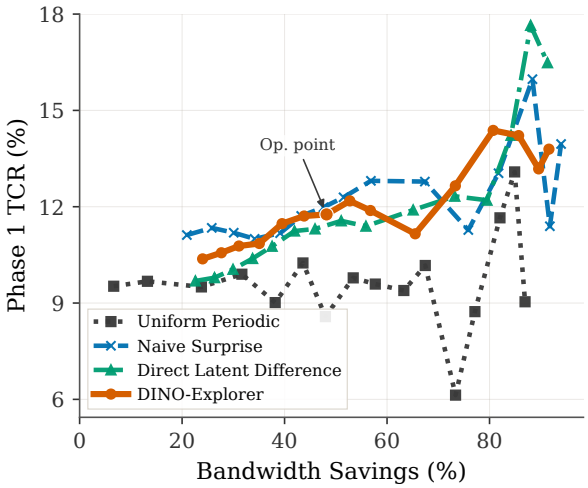
Table II reports the fixed-threshold replay comparison at the operating point highlighted in Fig. 4. DINO-Explorer achieves the best validated peak F1 while also providing the strongest bandwidth saving among the compared methods. Its Phase 1 SPCR, Phase 1 TCR, and LER remain close to the strongest diagnostic baselines.



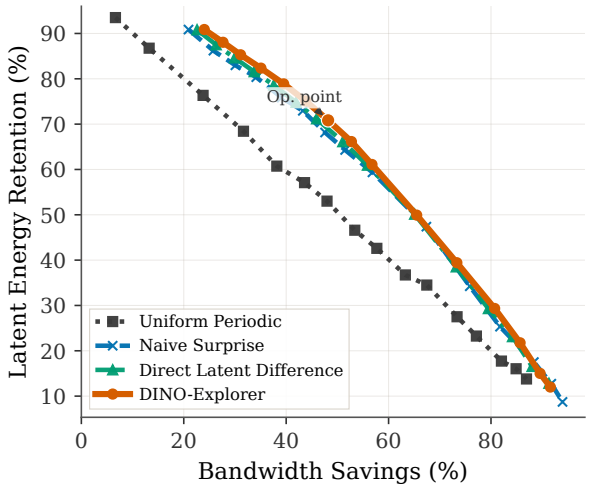
(a) Validated peak F1 versus telemetry budget



(b) Phase 1 SPCR recall-side retention versus telemetry budget



(c) Phase 1 TCR precision-side confirmation versus telemetry budget



(d) DINOv3-latent change energy retention versus telemetry budget

Fig. 4. Ablation benchmark for the proposed DINO-Explorer under matched telemetry budgets. Sweeping α avoids tying the comparison to a single downstream event-extraction hyperparameter and shows the ablation benchmark across extraction sensitivities. Each panel is a budget-controlled frontier: at the same bandwidth saving, a higher curve indicates stronger proposal quality or better preservation of the corresponding diagnostic. The sweep compares DINO-Explorer with the uncompensated surprise signal, Direct $\Delta \mathbf{z}_t$, and uniform replay, showing that the ego-motion-compensated surprise signal improves human-reviewer-facing proposal quality while preserving the Phase 1 recall, proposal-confirmation, and DINO-latent semantic-change diagnostics needed for underwater triage.

TABLE III
RUNTIME CONTEXT FOR ASYNCHRONOUS MISSION-TIME TRIAGE ON AN RTX A6000.

Metric summary	Value
Avg / median / p95 latency (ms)	227.2 / 228.7 / 272.9
Average FPS	4.40

D. Runtime and deployment context

We include runtime as deployment context for asynchronous triage and as a target for future onboard optimization.

Table III summarizes latency and throughput from the current inference logs.

VII. Discussion and Conclusion

We present DINO-Explorer, a motion-aware semantic predictive-coding framework that converts foundation-model latent prediction error into a continuous surprise signal for underwater robot attention in visually degraded marine environments. The evaluation on a representative downstream telemetry task supports three design claims. First, semantic prediction mismatch provides a useful abstraction for identifying scene-level novelty beyond low-level visual fluctuation. Second, conditioning prediction on ego-motion improves the separation between self-induced visual change and externally meaningful novelty. Third, the resulting surprise signal can serve as a decision variable for downstream robot-facing interfaces, as demonstrated here through event triage and

selective telemetry.

A. Limitations and Future Directions

The current implementation leaves two concrete extension points. First, the inference package is validated as an asynchronous server-side pipeline on the RTX A6000; turning DINO-Explorer into a real-time onboard module will require compressing or distilling the frozen DINOv3 encoder and RAFT motion estimator and scheduling the recurrent predictor under robot compute budgets. Second, the present representation applies global average pooling over DINO latent before predicting surprise. This keeps the signal compact, but it loses local surprise detail: when surprise signal spikes, the system cannot yet identify which patch, object, or local visual transition in the frame primarily drove that surprise. Future versions should preserve patch-level or multi-scale latent maps, compute localized surprise heatmaps to provide visual cues that can indicate where in the view made the robot surprised.

These future paths keeps the current contribution centered on a measured surprise signal while making its next role explicit: from asynchronous event triage, to active exploration control, and eventually to curiosity-driven semantic world-model refinement in long-horizon robot learning.

ACKNOWLEDGMENTS

This work was funded by the German Federal Ministry for the Environment, Climate Action, Nature Conversation and Nuclear Safety (BMUKN) supported by the ZUG under grants 67KIA4036A and 67KIA4036C, and partially supported by the German Federal Ministry of Research, Technology and Space (BMFTR) under the Robotics Institute Germany (RIG) under grant 16ME1010. The authors would like to thank Nael Jaber and Yi-Ling Liu for their valuable feedback and discussion on this manuscript.

REFERENCES

[1] Y. R. Petillot, G. Antonelli, G. Casalino, and F. Ferreira, "Underwater robots: From remotely operated vehicles to intervention-autonomous underwater vehicles," *IEEE Robotics & Automation Magazine*, vol. 26, no. 2, pp. 94–101, 2019.

[2] S. P. González-Sabbagh and A. Robles-Kelly, "A survey on underwater computer vision," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–39, 2023.

[3] A. F. Opdal, C. Lindemann, and D. L. Aksnes, "Centennial decline in north sea water clarity causes strong delay in phytoplankton bloom timing," *Global Change Biology*, vol. 25, no. 11, pp. 3946–3953, 2019.

[4] R. J. Wilson and M. R. Heath, "Increasing turbidity in the north sea during the 20th century due to changing wave climate," *Ocean Science*, vol. 15, no. 6, pp. 1615–1625, 2019.

[5] Y. Zhou, B. Li, J. Wang, E. Rocco, and Q. Meng, "Discovering unknowns: Context-enhanced anomaly detection for curiosity-driven autonomous underwater exploration," *Pattern Recognition*, vol. 131, p. 108860, 2022.

[6] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *European Conference on Computer Vision*, pp. 402–419, Springer, 2020.

[7] M. Ferrera, J. Moras, P. Trouvé-Peloux, and V. Creuze, "Real-time monocular visual odometry for turbid and dynamic underwater environments," *Sensors*, vol. 19, no. 3, p. 687, 2019.

[8] R. P. N. Rao and D. H. Ballard, "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects," *Nature Neuroscience*, vol. 2, no. 1, pp. 79–87, 1999.

[9] K. Friston, "A theory of cortical responses," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 360, no. 1456, pp. 815–836, 2005.

[10] E. von Holst and H. Mittelstaedt, "Das Reafferenzprinzip," *Naturwissenschaften*, vol. 37, no. 20, pp. 464–476, 1950.

[11] T. B. Crapse and M. A. Sommer, "Corollary discharge across the animal kingdom," *Nature Reviews Neuroscience*, vol. 9, no. 8, pp. 587–600, 2008.

[12] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, F. Massa, D. Haziza, L. Wehrstedt, J. Wang, T. Darcet, T. Moutakanni, L. Sentana, C. Roberts, A. Vedaldi, J. Tolan, J. Brandt, C. Couprie, J. Mairal, H. Jégou, P. Labatut, and P. Bojanowski, "Dinov3," 2025.

[13] R. Bajcsy, "Active perception," *Proceedings of the IEEE*, vol. 76, no. 8, pp. 966–1005, 1988.

[14] R. Bajcsy, Y. Aloimonos, and J. K. Tsotsos, "Revisiting active perception," *Autonomous Robots*, vol. 42, no. 2, pp. 177–196, 2018.

[15] E. Chen, T. Manderson, N. Karapetyan, P. Edmunds, N. Roy, and Y. Girdhar, "Autonomous search for sparsely distributed visual phenomena through environmental context modeling," *arXiv preprint arXiv:2603.10174*, 2026.

[16] E. C. Orenstein, B. Woodward, L. Lundsten, K. Barnard, B. Schlining, and K. Katija, "Assisting human annotation of marine images with foundation models," *Frontiers in Marine Science*, vol. 12, p. 1469396, 2025.

[17] Z. Chen, C. Zhang, H. Fang, and R. Cong, "Empowering dino representations for underwater instance segmentation via aligner and prompter," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, pp. 3201–3209, 2026.

[18] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell, "Curiosity-driven exploration by self-supervised prediction," in *Proceedings of the 34th International Conference on Machine Learning* (D. Precup and Y. W. Teh, eds.), vol. 70 of *Proceedings of Machine Learning Research*, pp. 2778–2787, PMLR, 2017.

[19] T. Taniguchi, S. Murata, M. Suzuki, D. Ognibene, P. Lanillos, E. Ugur, and G. Pezzulo, "World models and predictive coding for cognitive and developmental robotics: frontiers and challenges," *Advanced Robotics*, vol. 37, no. 13, pp. 780–806, 2023.

[20] D. Ha and J. Schmidhuber, "World models," 2018.

[21] D. Hafner, T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson, "Learning latent dynamics for planning from pixels," in *Proceedings of the 36th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.), vol. 97 of *Proceedings of Machine Learning Research*, pp. 2555–2565, PMLR, 2019.

[22] P. Wu, A. Escontrela, D. Hafner, P. Abbeel, and K. Goldberg, "Daydreamer: World models for physical robot learning," in *Proceedings of the 6th Conference on Robot Learning* (K. Liu, D. Kulic, and J. Ichnowski, eds.), vol. 205 of *Proceedings of Machine Learning Research*, pp. 2226–2240, PMLR, 2023.

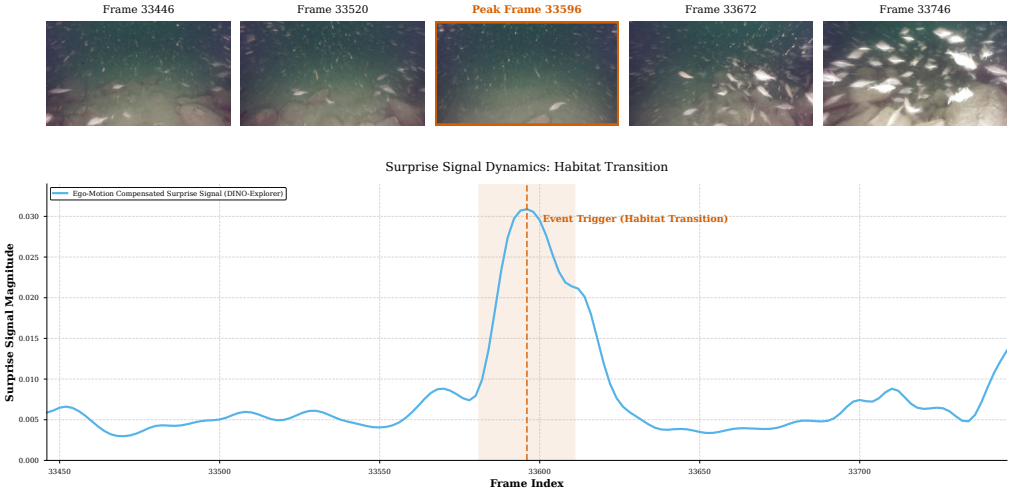
[23] G. Zhou, H. Pan, Y. LeCun, and L. Pinto, "Dino-wm: World models on pre-trained visual features enable zero-shot planning," 2025.

[24] S. E. A. Pineda-Metz, "Master tracks in different resolutions of HEINCKE cruise HE663, Bremerhaven - Bremerhaven, 2025-06-17 - 2025-07-01," 2025.

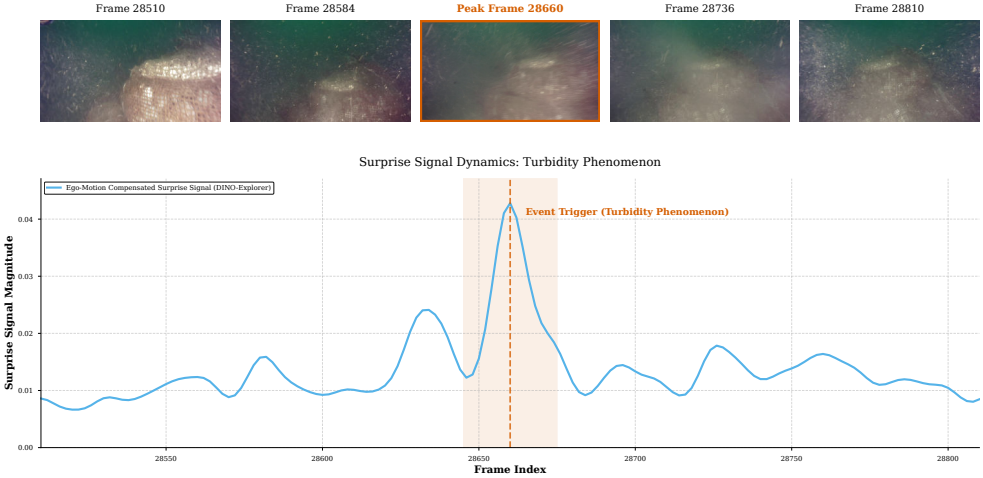
[25] Blue Robotics, "BlueROV2 (BROV2) Datasheet." https://bluerobotics.com/wp-content/uploads/2025/04/BROV2-DATASHEET.pdf, 2025. Accessed: 2026-03-18.

APPENDIX A
QUALITATIVE EXAMPLES OF SURPRISE EVENTS

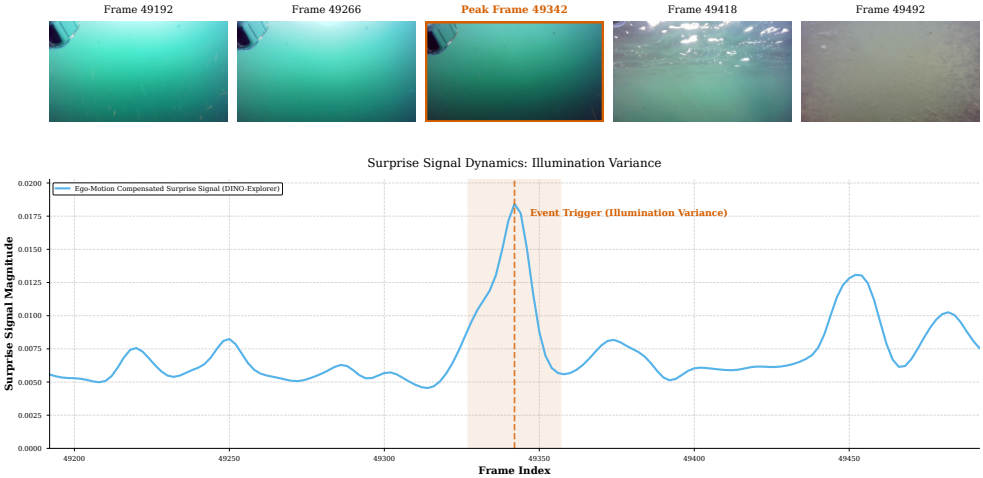
This appendix provides qualitative examples for three non-biological surprise categories: habitat transitions, turbidity bursts, and illumination changes. Each case shows five frames with the compensated surprise signal ($\mathcal{S}_{comp}$), illustrating low response to predictable underwater noise and strong peaks for event-level changes such as reef transition (Fig. 5a), sediment plume (Fig. 5b), and illumination shift (Fig. 5c).



(a) Habitat transition: semantic surprise rises as sandy substrate gives way to reef structure.



(b) Turbidity phenomenon: a sediment plume abruptly changes scene clarity.



(c) Illumination variance: ambient light or camera exposure shifts the global appearance.

Fig. 5. Qualitative surprise-event examples. Rows show the compensated surprise trace for a habitat transition, turbidity plume, and illumination shift, illustrating low response to predictable underwater noise and peaks on event-level semantic changes.