Real-Time Newsworthiness-Driven Journalist Robot: Perception-to-Publication with Multimodal Synchronization

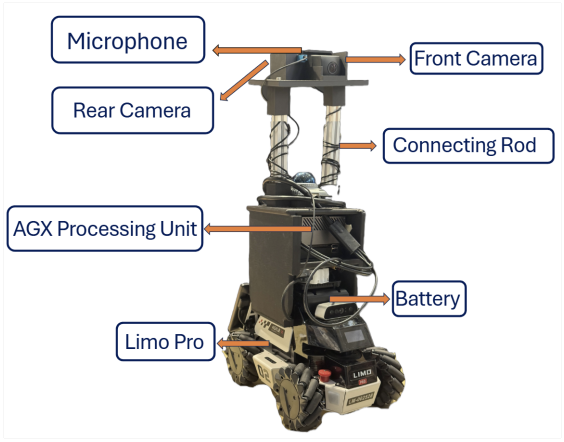
Anonymous Authors

***Abstract*— News organizations face growing coverage demands while the professional journalist workforce and on-site resources have not kept pace, creating event–reporter mismatches, delayed reporting, and undercoverage of events. However, existing automated journalism primarily automates content generation from prerecorded or web data, but it lacks a connection to the data collection process; as a result, systems still depend on human operators to capture suitable footage and to decide what to record. This separation between capture and generation delays coverage and lengthens the capture-to-publication loop, while continuing to rely on human operators and therefore not alleviating the limited journalist workforce. This paper proposes an autonomous journalist robot which integrates realtime newsworthiness assessment with navigation and a synchronized multimodal pipeline, ensuring that capture decisions are directly guided by editorial requirements. The system fuses dual-camera video and audio, performs semantic selection, generates grounded articles, and evaluates them automatically. Experimental validation uses three quantitative views: navigation alignment with a human operator, single vs multi-frame generation quality, and the article quality of the autonomous system. We measure navigation alignment at 94% (alignment with human operator), multi-frame article quality at +23.3% relative improvement over single-frame, and end-to-end article quality at +11.3% over a static recorded baseline.**

I. INTRODUCTION

Autonomous journalist robots are important for timely, wide-reaching access to trustworthy information. They can operate where human reporters face safety risks, geographic constraints, or resource limitations, reducing the time from event occurrence to public dissemination and improving situational awareness in fast-moving contexts such as disasters, protests, and emergencies. Beyond journalism, the same capabilities support industrial monitoring, environmental observation, and public safety applications where rapid, reliable information gathering is essential. Such systems therefore help alleviate newsroom resource shortages [1] while advancing equitable, real-time access to verified information in an increasingly complex world.

Automated journalism has become increasingly capable of producing coherent and informative content. However, current systems remain hindered by a disconnect between data collection and article generation [2]. Many frameworks rely on prerecorded or web-based datasets and require human operators to capture suitable footage and decide what to record [3], [4]. This separation prevents real-time adaptation, leading to missed vantage points, delayed responses, and continued dependence on limited human resources. In addition, multimodal pipelines often struggle to fuse audio and video coherently, relying heavily on visual streams while neglecting synchronized grounding, which weakens factual consistency [5], [6]. These limitations motivate autonomous journalist robots that integrate sensing, navigation, and generation into a unified system.



(a) The robot hardware platform.

(b) Example scene perception.

Headline: New Robot Takes Shape in Tohoku University Lab

Article Text: SENDAI – In a university robotics workshop, two researchers work side-by-side at a central bench, assembling a tracked mobile robot. The robot sits atop the table with its black tread and sensor mounts clearly visible as one researcher adjusts wiring while the other reviews components. Around them, shelves of labeled bins, tools, and test equipment frame the scene, highlighting an active lab environment. The focus remains on the robot under construction, but the organized background underscores ongoing development and the team's hands-on, technical workflow.

(c) The resulting auto-generated news article.

Fig. 1: An overview of our autonomous journalism system, showing (a) the mobile platform, (b) a perceived scene, and (c) the final generated article.

Our system provides a complete perception to publication pipeline, as illustrated in Fig. 1.

In this paper, we propose a fully autonomous journalist robot that tightly couples real-time data collection with

article generation. The system employs vision–language models for scene understanding and robot control, enabling autonomous movement and newsworthiness-driven data collection from dual cameras that provide 360° coverage. Multimodal data are processed via visual analysis and semantic selection, and a language model synthesizes grounded articles based on synchronized frame descriptions and audio transcripts. Audio–video synchronization ensures factual narratives, and a closed-loop architecture performs on-the-fly evaluation that feeds back into selection and motion policies, operating without human intervention to deliver low-latency, autonomous journalism from sensing to publication.

Our contributions are: (i) newsworthiness-driven autonomous data collection that dynamically controls robot motion and capture decisions to improve positioning and selectivity, with feedback from the journalism process used to refine pose and coverage quality; (ii) a unified real-time journalism pipeline that couples data collection with content production to reduce capture-to-publication latency; (iii) synchronized multimodal processing that aligns audio–visual evidence and generates grounded journalistic content from frame descriptions and transcripts; and (iv) a comprehensive multimodal evaluation framework with 10 specialized spatial–temporal–semantic metrics, including enhanced factual verification, video relevance scoring, and journalistic quality measures not available in existing frameworks.

II. RELATED WORK

The related work in autonomous journalism robotics can be categorized into three main areas:

A. Data Processing and News Content Generation

Data-to-text systems establish the foundation for automated journalism and motivate our focus on integrating collection with generation. Foundational works convert structured data into narrative content; Diakopoulos [3] and Graefe [2] analyze how rule-based NLG turns financial reports, election results, and sports statistics into articles. Modern implementations include Reuters Tracer [4] for social media streams, Xiaomingbot [7] for multilingual news from structured databases, and BLAB Reporter [8] for environmental journalism from sensor data. Notably, these systems primarily process non-visual sources (numeric/text/metadata) rather than multimedia, and evaluation frameworks such as SelfCheckGPT [9], FactScore [10], QAFactEval [11], and SummEval [12] target factuality and quality. However, reliance on pre-structured sources leaves a disconnect from live collection, creating timeliness and adaptability gaps in dynamic news environments.

Multimodal generation research frames our use of extractive and abstractive methods for journalistic content. Core approaches include extractive selection of key segments/frames and abstractive narrative generation from multimedia [13]. VLMs such as LLaVA-NeXT-Interleave [14], CLIP [15], and BLIP-2 [16], combined with LLMs like DeepSeek R1, have advanced multimodal understanding for journalism. The MAST framework [5] demonstrates trimodal hierarchical attention, yet most work remains dual-modal rather than full audio-visual-text fusion. While MHMS [6] enables cross-domain alignment, end-to-end systems that simultaneously process audio, visual, and textual modalities for journalistic generation remain scarce, leaving integrated multimodal processing underexplored.

B. Data Collection

Data collection methods motivate our fully autonomous approach. Traditional journalism relies on human crews for gathering and curating information, while recent robotics introduce mobile and telepresence systems to capture visual and audio data. Examples include social robots integrated with LLMs for conversational reporting [17], demonstrating AI-enhanced reporters. However, most systems remain teleoperated—requiring continuous human supervision and following predefined routes or static setups [18]–[20]—which limits scalability and induces operator fatigue and latency. A key gap is the absence of complete autonomy, where capture choices are not optimized for journalistic salience [21], [22]. Our work closes this gap with newsworthiness-driven control and closed-loop multimodal collection, removing human dependency while achieving real-time editorial optimization.

C. Robot Embodiment

Embodiment research underpins our navigation and decision-making design. Surveys synthesize how morphology, perception, action, and learning jointly enable agents to act in dynamic environments [23]. In particular, VLN has progressed with hierarchical memory and dual-scale graph reasoning for instruction-following in complex scenes [24], and language-grounded control has been demonstrated by grounding high-level semantics into affordance-aware policies for real robots [25]. Robotics transformers, exemplified by RT-1 [26], further integrate large-scale language models with robotic control for end-to-end manipulation and navigation. These strands show that vision–language models can improve autonomous movement and closed-loop decisions. However, no prior work applies embodied navigation to journalism or implements newsworthiness-driven navigation, underscoring our novelty: directly coupling editorial newsworthiness with embodied control to guide real-time data acquisition.

Addressing these gaps, our system (1) closes the collection–generation disconnect by coupling capture policy to on-the-fly editorial signals, avoiding reliance on pre-structured inputs and post-hoc processing [2]–[4], [7], [8]; (2) replaces teleoperation and predefined routes with autonomous, newsworthiness-driven navigation and synchronized audio–visual capture, removing the human-in-the-loop bottleneck observed in recent mobile/telepresence data-collection frameworks [18]–[20]; (3) overcomes multimodal fusion limits by aligning acquisition with editorial intent during capture rather than relying solely on post-hoc summarization [5], [6], [13]; and (4) operationalizes embodied navigation for journalism with

policy decisions directly driven by a real-time newsworthiness score—capabilities absent from prior embodied AI work [23]–[25]. Collectively, this yields a fully autonomous, end-to-end system that improves coverage fidelity, reduces capture-to-publication latency, and eliminates operator workload while producing intent-aligned outputs.

III. METHODOLOGY

We summarize the overall workflow in Fig. 2; the following subsections explain these processes.

A. Detection, Perception, and Newsworthiness Assessment

To extract scenes for journalism, scene recognition is performed by conditioning the base model on mission-specific background context (a concise brief describing subjects of interest, safety constraints, and editorial priorities), mimicking the workflow of human journalists. Foundation vision–language models (VLMs) such as CLIP [15], Flamingo [27], LLaVA [28], and VideoLLaMA3 [29] are trained on large, heterogeneous web-scale corpora, yielding broad, general-purpose perceptual priors. In unconstrained, dynamic reporting settings, we observed that using these models "as is" produces diffuse and temporally inconsistent scene interpretations that are insufficient for reliable editorial decisions. To specialize perception for journalism, we condition the model on a mission-specific *background context*—a compact brief describing subjects of interest, safety constraints, and editorial priorities—analogous to how human journalists operate. Concretely, we inject this context into the multimodal prompt so that perception and subsequent movement choices are explicitly evaluated against newsworthiness criteria rather than generic salience.

The background context (mission objectives, priorities, editorial criteria) is injected into the prompt to specialize perception for journalism tasks. The prompt uses structured elements—system role, confidence calibration thresholds, and a newsworthiness rubric—to guide decision reliability; see the excerpt below.

To enable responsive control in dynamic scenes, we segment real-time video into short ($\sim 2\,$s) windows, which shortens the perception–action cycle and mitigates VideoLLaMA3's non–real-time-optimized architecture [29]. Each window is then evaluated using: (i) a dynamic background context (mission brief, priorities, and newsworthiness criteria) that can be updated externally, (ii) a stable instruction-following prompt (see the prompt excerpt box) that defines the output schema and robot control interface, and (iii) a concise summary of recent commands to reduce repetition and encourage exploratory coverage. VideoLLaMA3 processes this combined input and returns a single, structured response that couples natural-language reasoning describing the scene and its relevance to the background context with a movement proposal, a newsworthiness score, a confidence score, and an explicit start/continue/stop trigger for recording and journalism pipeline activation [29]. Once the journalism pipeline is started, the robot continues exploratory navigation while awaiting feedback on captured data quality and any movement adjustments recommended by the journalism process to improve coverage.

> Prompt excerpt (controller): *System role:* Intelligent exploratory journalist robot controller. *The mission background (subjects, priorities, safety) is injected verbatim; all decisions must be grounded in the current video and the brief.* Confidence calibration*: high (0.8–1.0), medium (0.6–0.8), low (0.4–0.6), very low (0.0–0.4) ⇒ speed/safety scaling.* Vision-first rules*: forward 0.6–0.8 m/s when targets/open path are visible; rotation 1.2–2.0 rad/s for scanning/reframing.* Safety*: never move forward if the center path appears blocked; maintain distance; use conservative speeds at low confidence.* Output instruction*: return concise natural-language reasoning and the movement commands (*`linear_x`*,* `angular_z`*) with a confidence value; omit any other text.*

Newsworthiness and control. We smooth the model's per-window newsworthiness estimate and form an overall score N_t as a weighted sum of three signals: smoothed newsworthiness, agreement with the mission brief, and optional novelty (weights sum to 1). Recording begins when $N_t \geq 0.7$ and confidence $c_t \geq 0.8$ and stops when both drop below 0.3 and 0.5, respectively (hysteresis). Motion is tied to N_t and path clearance: when the path is clear and N_t is high the robot moves forward; otherwise it rotates to reframe, with speeds scaled down under low confidence. Safety reflexes (obstruction turns), hysteresis, and emergency stops are always active. Decision policy: start the journalism pipeline when $N_t \geq \tau_w$ and $c_t \geq \tau_c$ with τ_w=0.7 and τ_c=0.8; stop when N_t<0.3 and c_t<0.5. Let $F_t \in \{0,1\}$ indicate a clear forward path and $d_t \in \{-1,+1\}$ the preferred turn direction. The commanded velocities are

$$v_t^{\text{lin}} = v_{\max} N_t F_t, \quad v_t^{\text{ang}} = \omega_{\max}(1-N_t)(1-F_t)d_t,$$

optionally scaled by confidence, $(v_t^{\text{lin}}, v_t^{\text{ang}}) \leftarrow \sigma(c_t)(v_t^{\text{lin}}, v_t^{\text{ang}})$ with $\sigma(c) \in \{0.4, 0.6, 0.8, 1.0\}$. The journalism pipeline is triggered only during high-N_t segments so that processed evidence aligns with key moments.

Rationale and sensitivity. Thresholds τ_w=0.7 and τ_c=0.8 were selected via a small validation sweep ($\tau_w \in \{0.6, 0.7, 0.8\}$, $\tau_c \in \{0.7, 0.8, 0.9\}$) on held-out runs to balance false starts (off-brief/low-confidence recording) against missed opportunities. Settings lower than 0.7/0.8 increased off-topic segments; higher values reduced recall of fleeting events.

B. Journalism Pipeline: Multimodal Data Processing and Content Generation

Once a newsworthiness event is triggered, the journalist pipeline begins continuous, real-time monitoring of the combined dual-camera feed (front + rear, providing 360° coverage) together with the live audio stream, and prepends a $\sim$10 s pre-event context window to preserve environmental history. The 10 s horizon was chosen as a practical compromise observed in pilot runs: shorter windows ($< 5\,s$) omitted setup cues relevant to our scenarios (e.g., subject handovers, equipment operation), while much longer windows increased latency and evidence retrieval cost without added benefits; 10 s matched the typical utterance/shot change scale we encountered. The first step is a fast quality and relevance check using a vision–language model (LLaVA) [28], [30].

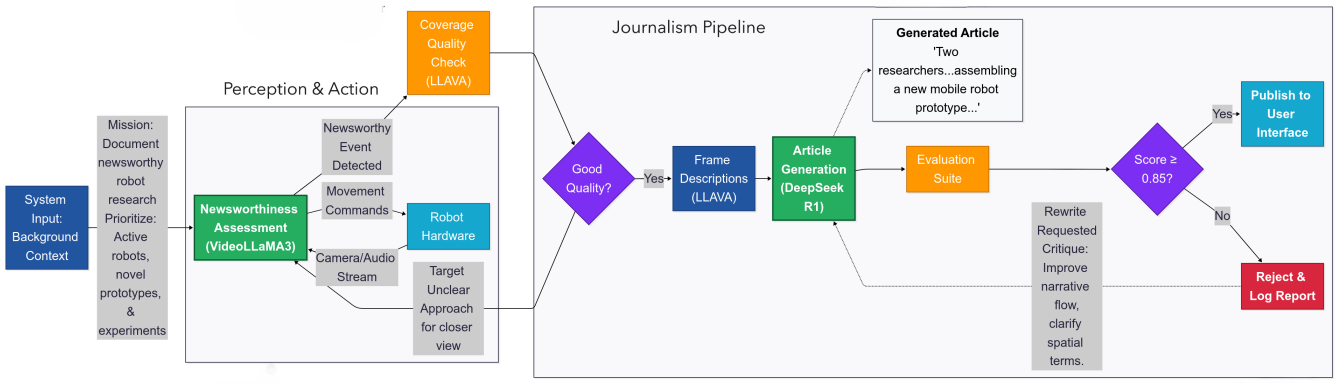


Fig. 2: The proposed autonomous journalism pipeline, showing the Perception & Action loop and the Journalism Pipeline.

TABLE I: Effect of Semantic Selection on Processing (Three Runs)

Experiment	Total Frames	Selected Frames	Processing Time (s)
Lab: Empty Desks	2500	600	136 (vs 294 w/o selection)
Lecture: Screen+Board	2500	480	128 (vs 286 w/o selection)
Workshop: Equipment	2500	410	122 (vs 272 w/o selection)

If the captured view is unclear, off-topic, or poorly framed, the pipeline issues structured feedback to the perception and newsworthiness loop with actionable guidance (e.g., rotate toward the subject, adjust distance/centering) to improve coverage before continued recording. If the data quality is sufficient, the pipeline proceeds as follows.

Semantic selection. Visual redundancy is reduced via semantic selection using CLIP (ViT-B/32) [15] to retain frames that best represent salient scene changes while discarding near-duplicates. This extractive step reduces processing time without sacrificing coverage quality by preserving the information needed for abstractive generation, consistent with extractive/abstractive video summarization practice [13]. Selected frames are grouped into compact batches to preserve short-range temporal coherence. Table I summarizes typical reduction and processing effects over three representative runs.

Multimodal description. For each batch of 8 frames, a multimodal VLM (LLaVA) produces concise, time-ordered visual descriptions grounded in the background context [28], [30], while the audio stream is transcribed by automatic speech recognition (Whisper) [31] and aligned on the same timeline. The resulting text units (visual descriptions and speech transcripts) are synchronized so that auditory events and visual observations are associated with the correct moments. To maintain a consistent narrative and improve temporal understanding, each new batch is processed together with the immediately preceding batch's visual descriptions, reducing hallucination and strengthening scene and environment continuity across batches.

Prompted synthesis and grounding. For article generation, the pipeline constructs a structured prompt with (i) the mission background context, (ii) the synchronized visual descriptions and transcripts, and (iii) a brief timeline of key events. A retrieval-augmented grounding step indexes the accumulated descriptions/transcripts to ensure that generation cites the most relevant evidence and so that information can be efficiently retrieved later to answer detailed queries. These grounded inputs are then synthesized by a reasoning-augmented language model (DeepSeek-R1) using a structured prompt that explicitly binds evidence segments to the requested headline, lead, and body, yielding a long-form article that follows the inverted-pyramid structure—leading with the most newsworthy information before elaborating supporting details [32]. In ablation runs that used a chronological, no-structure prompt, drafts exhibited lower lead informativeness and required more revision cycles to reach the 0.85 acceptance threshold; the inverted-pyramid schema produced more first-pass acceptances and higher temporal/spatial consistency under the same evaluator. The article is output in text and audio formats and is then passed to the evaluation algorithms; the system awaits feedback to guide any further refinements.

TABLE II: Evaluation Dimensions and Models

Dim.	Purpose	Model/Signal
Visual–Text Rel.	Evidence–text match	CLIP sim./retrieval
Multimodal Fact.	Grounding; claim entailment	BLIP/VQA; RoBERTa–MNLI
Sem./Temp./Spat. Consistency	Across frames + descriptions	LLaVA-Interleave eval.
Language Quality	Coherence; grammar/style; structure	SBERT; Language-Tool; structure

C. Evaluation and Acceptance

Given an article draft with synchronized frame descriptions, selected keyframes, and the audio transcript, the evaluation framework (Sec. III-C) runs as a separate process and scores outputs across the defined dimensions; the metrics and signals are summarized in the Evaluation Dimensions table earlier in the paper.

Evaluation aggregates complementary signals to

gate acceptance. Article sentences are embedded with Sentence-BERT [33] to measure intra-article coherence; keyframes and batch descriptions are scored for semantic/visual/temporal agreement via LLaVA-NeXT-Interleave; claim–evidence pairs are tested with RoBERTa-MNLI [34]; and CLIP provides sentence–frame relevance and retrieval-style recall. Scores are aggregated with configuration weights to produce an overall evaluation score in $[0, 1]$. Decision policy: if the overall score is less than 0.85 across the active metrics, the system routes the package back to article generation with a structured evaluation report; otherwise (≥ 0.85), the article is accepted post-evaluation and grounding for publication.

Upon acceptance, the article is served through a chat interface backed by retrieval-augmented generation over the recorded evidence.

IV. Experimental and Evaluation Setup

A. Hardware Setup

As illustrated in Fig. 1(a), we mount two Tier IV C2 cameras—one capturing the scene in front of the robot and one capturing the scene behind it—to deliver continuous 360° visual coverage at newsroom-appropriate quality. A ReSpeaker Mic Array captures scene audio and direction of arrival for spatial cues. All sensors connect to an on-board AGX processing unit that handles frame/audio ingest, light pre-processing, and ROS 2 streaming to a remote GPU server for perception, newsworthiness assessment, and content generation. Motion is executed by a Limo Pro mobile base (Jetson Nano controller) that receives velocity commands over ROS 2.

Compute split and ROS 2 organization. On-board AGX nodes (ROS 2) handle sensing (drivers, time-sync, compression, motion I/O), while a remote GPU runs perception/newsworthiness and a separate service runs journalism/evaluation; three logical nodes cover newsworthiness, motion control, and journalism/evaluation. End-to-end control latency is 130–170 ms over campus Wi-Fi (real-time navigation), and article-pipeline latency scales with selected frames (Table I).

Two custom 3D-printed enclosures organize the system: a lower chassis case (AGX, batteries, camera interfaces) mounted to the Limo Pro deck, and an upper sensor mast case (dual cameras, microphone). We used a 1 m sensor rod (mast) to approximate person-eye viewpoint and reduce occlusions in crowded scenes. Components are integrated through ROS 2: cameras and microphone publish streams to the AGX, which relays to the GPU server; inference results return as twist commands for execution on the Limo Pro.

B. Experiments and Evaluation

To evaluate this system, we assess multiple stages: the newsworthiness navigation control, the generated article quality, and the full end-to-end integration. The following subsections describe the experimental setups used for each stage.

1) Newsworthiness-Driven Navigation: We evaluate navigation across three environments (Research Lab, Research Workshop, Lecture Room) using four approaches. *Procedure*: each run executes a fixed budget of 70 decisions; a "decision" is one discrete control proposal from the $2\,\mathrm{s}$ newsworthiness window that, after safety gating, issues a forward or turn command. Thus each run spans $140\,\mathrm{s}$ of closed-loop operation. *Metrics*: (1) decision alignment with the human baseline (percentage of executed actions that match the human's forward/turn choice at the same decision index), (2) coverage quality (percentage of time with a clear view of the scene subject), (3) total distance traveled (meters). *Human-operator comparison*: in each environment we run four experiments under identical conditions: (1) Teleoperation baseline—a human operator drives the robot to collect data and all movement is recorded; (2) Autonomous without background—the controller runs using only a generic journalist-role prompt, without a mission background context; (3) Autonomous with background only—the controller is conditioned on a mission-specific background context; (4) Autonomous with background and journalism feedback—the controller also receives coverage feedback from the journalist pipeline. We then compare the movement decisions across these experiments to assess alignment with the human operator. Results: Table III.

2) Effect of Background-Context Variation on Navigational Decisions: We repeat autonomous runs in the same environment and setting with two different background contexts, and replicate this pair across three different environments. This isolates how the mission brief steers perception and movement choices under otherwise identical conditions. Results: Table IV and Fig. 3.

3) Single-Frame vs. Multi-Frame Processing: Initially, we used a single-frame generator, which accepts one frame at a time; timelines and cross-frame relevance were harder to maintain and sometimes yielded brittle descriptions. We then switched to an LLaVA-based formulation that accepts multiple frames per query [28], [30]. We tested both approaches in the same settings and evaluated outputs with the above algorithm to measure the effectiveness of multi-frame processing. Results: Table V.

4) Generated Article Quality: End-to-End System Evaluation: We conduct experiments with six different scenarios in three environments and, using the previously described evaluation algorithm, assess the generated articles to measure overall quality. To evaluate the functionality and effectiveness of the full system in bridging the gap between data collection and content generation, we test end-to-end processing and article creation using recorded data and the newsworthiness-driven navigation approach in the same environment settings. This assesses how effectively the feedback loop between journalism generation and newsworthiness assessment improves coverage quality and clarity, capturing important events without missing key scenes or producing low-quality footage. Results: Table VI.

TABLE III: Multi-Metric Performance Comparison of Navigation Controllers

Approach	Total	Aligned (%)	Coverage (%)	Dist. (m)
Human Baseline	70	100%	95%	15.2
Generic Prompt	70	42%	38%	21.5
With Context	70	82%	85%	16.5
Context + Feedback	70	94%	93%	15.5

TABLE IV: Analysis of Behavioral Divergence

Environment	Context A	Context B	Shared (70)	Divergent (70)	Divergence (%)
Research Lab	Focus on People Working	Focus on Empty Desks	7	63	90%
Research Workshop	Capture Robotics Advancements	Focus on Industrial Equipment	10	60	86%
Lecture Room	Focus on Screen and Board	Focus on the Presenter	6	64	91%

V. RESULTS

This section reports the outcomes for the three evaluation tracks introduced in Section IV: (1) newsworthiness-driven navigation, (2) single- vs multi-frame article generation, and (3) end-to-end article quality. Experimental procedures and metrics follow Section IV-B and are summarized here only where needed for clarity.

A. Newsworthiness-Driven Navigation

The aggregated comparison across the four approaches under a fixed budget of 70 decisions is presented in Table III.

For the fixed budget of 70 decisions, the generic-prompt controller aligned with the human operator on 42% of decisions; conditioning on background context increased alignment to 82%; adding journalism feedback yielded 94% alignment with 93% coverage quality.

B. Effect of Background-Context Variation on Navigational Decisions

As defined in Section IV-B, we assess Behavioral Divergence (%) between two autonomous runs per environment using distinct mission briefs (70 decisions per run). Results are in Table IV, with a qualitative example in Fig. 3.

The data presented in Table IV shows a high degree of behavioral divergence in all tested environments when the mission context was altered. The average divergence across the three scenarios was 89%. In the Lecture Room experiment, a comparison between the two runs revealed that only 6 of the 70 movement decisions were shared, corresponding to a 91% divergence in navigational strategy. Similarly, the Research Lab and Research Workshop environments showed high divergence rates of 90% and 86%, respectively.

C. Single-Frame vs. Multi-Frame Processing

The single-frame baseline (Llama 3.2) processes each piece of visual evidence in isolation, while the multi-frame model (LLaVA) can reason across a sequence of images. We expected that this would improve narrative and temporal consistency. The aggregated results (averaged over three runs per method) are presented in Table V; the multi-frame processing approach achieved a higher Overall Score (+23.3

TABLE V: Evaluation of Single-Frame vs. Multi-Frame Approaches

Evaluation Metric	Single-Frame (Llama 3.2)	Multi-Frame (LLaVA)	Performance Gain
Multimodal Factuality	0.75	0.89	+18.7%
Semantic Alignment	0.80	0.91	+13.8%
Spatial Reasoning	0.65	0.88	+35.4%
Temporal Consistency	0.58	0.90	+55.2%
Evidence Quality	0.78	0.89	+14.1%
Video Relevance	0.77	0.87	+13.0%
Overall Score	0.722	0.890	+23.3%

D. Generated Article Quality: End-to-End System Evaluation

We compare two data-collection modalities across three scenarios using the same metrics as above. Table VI reports per-scenario scores in [0,1] for Multimodal Factuality, Semantic Alignment, Spatial Reasoning, Temporal Consistency, Evidence Quality, Video Relevance, and an Overall Score. The 'After Eval.' column shows post-evaluation values: packages with Overall < 0.85 are returned for strengthening; accepted packages meet or exceed 0.85.

Across the three representative experiments, the autonomous system achieved higher Overall Scores than the static recorded baseline in every case. Averaged across experiments, Overall Score increased from 0.781 (static recorded) to 0.869 (autonomous), a +11.3% relative gain. Per-scenario Overall Scores improved as follows: People Working 0.780→0.863, Presenter 0.718→0.812, Robotics 0.705→0.784. After evaluation, static runs were revised to 0.857/0.854/0.851 to reach the 0.85 threshold, while autonomous outputs usually did not require revision.

VI. DISCUSSION

A. Role of Context and Feedback in Navigation

Our first set of experiments (Table III) shows the impact of contextual guidance on autonomous decision-making. The baseline controller, operating with only a generic prompt, achieved 42% alignment with the human operator. This indicates that without a clear mission, a VLM can be distracted, resulting in hesitant navigation and lower coverage quality.

The introduction of a mission-specific background context increased decision alignment to 82%. This indicates that the system's behavior is not reactive to visual saliency alone but is guided by a high-level, semantic understanding of its journalistic goals. This is further supported by the context-variation experiments (Table IV), where changing the mission brief induced a nearly 90% divergence in navigational strategy in identical environments. The robot did not just film "activity"; it selectively filmed activity as defined by its mission.

We observed the largest measured effect with the journalism feedback loop. The full system achieved 94% decision alignment, approaching the human operator. While a background context provides the strategic "what to film," the feedback loop provides operational guidance on "how to adjust framing." It addressed issues such as occlusion and framing, reflected in Coverage Quality (93%) and article scores that averaged 0.869 overall. These results are

TABLE VI: Comparative Evaluation of Article Quality (Static Recorded Data vs. Full Autonomous System)

Scenario	Modality	Multimodal Factuality	Semantic Alignment	Spatial Reasoning	Temporal Consistency	Evidence Quality	Video Relevance	Overall Score	After Eval. (≥ 0.85)
Lab: People Working	Static Recorded	0.80	0.82	0.75	0.74	0.79	0.78	0.780	0.857
	Autonomous	0.88	0.90	0.84	0.83	0.87	0.86	0.863	0.863
Lecture: Presenter	Static Recorded	0.74	0.76	0.69	0.68	0.72	0.71	0.718	0.854
	Autonomous	0.83	0.85	0.79	0.78	0.82	0.80	0.812	0.852
Workshop: Robotics	Static Recorded	0.72	0.74	0.67	0.66	0.70	0.69	0.705	0.851
	Autonomous	0.81	0.82	0.75	0.76	0.79	0.77	0.784	0.851
Average Score	Static Recorded	0.803	0.818	0.755	0.745	0.790	0.780	0.781	0.864
	Autonomous	0.885	0.900	0.848	0.842	0.875	0.862	0.869	0.869

consistent with a relationship between intelligent navigation, the quality of collected data, and the quality of the final journalistic output.

B. The Necessity of Multi-Frame Processing for Coherent Narratives

The comparative evaluation of single-frame versus multi-frame processing (Table V) provides a clear justification for our architectural choice. The single-frame model, despite its power, struggled to create narratives, as evidenced by its particularly low scores in Temporal Consistency (0.58) and Spatial Reasoning (0.65). It perceives the world as a series of disconnected snapshots.

The multi-frame LLaVA-based model, in contrast, demonstrated an improved ability to understand the flow of events over time. The 55.2% performance gain in Temporal Consistency is the most telling metric. By reasoning across a sequence of images, the model can understand cause and effect, track subjects, and build a coherent story. This ability to maintain a consistent narrative is relevant to journalistic quality and supports the use of video-centric foundation models for this task.

C. The Value of Embodiment: Autonomous vs. Static Data Collection

The end-to-end system evaluation (Table VI) indicates that embodiment improves article quality relative to static capture; averaged across the three experiments, Overall Score increased from 0.781 (static) to 0.869 (autonomous). Rather than reiterating the per-scenario values, we interpret the pattern: autonomy reduces occlusions, improves framing, and maintains subject continuity. These changes are reflected as gains in spatial/temporal reasoning and evidence quality, which in turn support higher semantic alignment and multi-modal factuality. The "After Eval." outcomes further suggest that static capture often lacks sufficient evidence on first pass, requiring revisions to reach acceptance, whereas autonomous capture typically meets the threshold without change. Collectively, the results support that active viewpoint selection and synchronized audio–visual evidence are central to producing publishable, grounded articles with fewer downstream edits.

D. Limitations and Future Work

While our results are promising, we acknowledge several limitations that provide clear avenues for future research. Our experiments were conducted in structured, indoor environments. The system's robustness in chaotic, uncontrolled outdoor settings remains to be tested. Furthermore, while our newsworthiness model is effective, it is guided by explicit context; future work could explore models capable of inferring more abstract or emergent newsworthiness without a detailed brief. Finally, our hybrid-computing model relies on a remote GPU server. Future advancements in on-board AI accelerators may enable the entire pipeline to be run directly on the robotic platform, achieving complete computational autonomy. Future research will also explore multi-robot collaboration, where a team of journalist robots could coordinate to cover large-scale events from multiple perspectives.

VII. CONCLUSION

This paper introduced a fully autonomous journalist robot designed to bridge the critical gap between data collection and content generation. By integrating real-time newsworthiness assessment, closed-loop navigational control, and a multi-modal generation pipeline, we have demonstrated a system that moves beyond the traditional, disconnected paradigm of automated journalism.

Our experimental results provide strong, quantitative evidence for the effectiveness of this integrated approach. We have shown that a navigation controller guided by mission context and journalism feedback can achieve 94% decision alignment with a human operator, and can completely alter its data collection strategy based on the mission brief. This intelligent navigation translated into a higher quality final product. The full autonomous system produced articles that scored, on average, 11.3% higher than those generated from static, pre-recorded data, indicating a relationship between the quality of embodied data collection and the quality of the final journalistic output.

In conclusion, this work represents a step towards the realization of autonomous journalistic agents. By endowing a robot with the ability to perceive, decide, and act based on a dynamic assessment of newsworthiness, we have created a system that is not a content generator alone but an active participant in the newsgathering process. Future work will focus on extending this framework to more complex, unstructured outdoor environments and exploring multi-robot collaboration, paving the way for a future where autonomous systems can provide safe, timely, and insightful coverage of events around the world.

REFERENCES

[1] P. R. Center, "U.s. newsroom employment has fallen 26% since 2008," 2021, accessed: 2025-09-07. [Online]. Available: https://www.pewresearch.org/short-reads/2021/07/13/u-s-newsroom-employment-has-fallen-26-since-2008/

[2] A. Graefe, "Guide to automated journalism," Tow Center for Digital Journalism, Tech. Rep., 2016.

[3] N. Diakopoulos, *Automating the News: How Algorithms are Rewriting the Media*. Harvard University Press, 2019.

[4] D. Tracy, C. Dyer, M. Goldberg, M. Magdon-Ismail, and F. Menczer, "Toward automated news production using large scale social media data," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1848–1858.

[5] A. Khullar and U. Arora, "Mast: Multimodal abstractive summarization with trimodal hierarchical attention," *arXiv preprint arXiv:2010.08021*, 2020. [Online]. Available: https://arxiv.org/abs/2010.08021

[6] J. Qiu, J. Zhu, M. Xu, F. Dernoncourt, T. Bui, Z. Wang, B. Li, D. Zhao, and H. Jin, "MHMS: Multimodal hierarchical multimedia summarization," *arXiv preprint arXiv:2204.03734*, 2022. [Online]. Available: https://arxiv.org/abs/2204.03734

[7] Y. Zhu, J. Pan, Y. Zhou, C. Yang, Z. Hu, Y. Jiang, T. Yang, B. Wang, B. Zhang, C. Wang, *et al.*, "Xiaomingbot: A multilingual robot news reporter," *arXiv preprint arXiv:2007.08005*, 2020. [Online]. Available: https://arxiv.org/abs/2007.08005

[8] H. Oliveira, M. Melo, N. Silva, V. Costa, T. Almeida, J. Carvalho, and A. Jorge, "Blab reporter: Automated journalism covering the blue amazon," *arXiv preprint arXiv:2210.06431*, 2022. [Online]. Available: https://arxiv.org/abs/2210.06431

[9] T. Fu, H. Peng, S. Yue, J. Wang, K. Chen, *et al.*, "Selfcheckgpt: Detecting large language model hallucinations with self-consistency," *arXiv preprint arXiv:2303.08896*, 2023. [Online]. Available: https://arxiv.org/abs/2303.08896

[10] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi, "Factscore: Fine-grained evaluation of factual precision in large language models," *arXiv preprint arXiv:2305.14251*, 2023. [Online]. Available: https://arxiv.org/abs/2305.14251

[11] O. Honovich, T. Scialom, O. Levy, T. Schick, and Y. Belinkov, "Qafacteval: Improved qa-based factual consistency evaluation for summarization," *arXiv preprint arXiv:2112.08542*, 2021. [Online]. Available: https://arxiv.org/abs/2112.08542

[12] P. Laban, P. Langlais, and G. Lapalme, "Summeval: Re-evaluating summarization evaluation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 646–660, 2020.

[13] T. Alaa, A. Mongy, A. Bakr, M. Diab, and W. Gomaa, "Video summarization techniques: A comprehensive review," *arXiv preprint arXiv:2410.04449*, 2024. [Online]. Available: https://arxiv.org/abs/2410.04449

[14] P. Zhang, Y. Li, *et al.*, "LLaVA-NeXT-Interleave: Tackling multi-image, video, and 3D in large multimodal models," *arXiv preprint arXiv:2407.07895*, 2024. [Online]. Available: https://arxiv.org/abs/2407.07895

[15] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[16] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023. [Online]. Available: https://arxiv.org/abs/2301.12597

[17] A. Hireche, A. N. Belkacem, S. Jamil, and C. Chen, "Newsgpt: Chatgpt integration for robot-reporter," *arXiv preprint arXiv:2311.06640*, 2023. [Online]. Available: https://arxiv.org/abs/2311.06640

[18] S. Huang, Y. Liao, S. Feng, S. Jiang, S. Liu, H. Li, *et al.*, "Adversarial data collection: Human-collaborative perturbations for efficient and robust robotic imitation learning," *arXiv preprint arXiv:2503.11646*, 2025. [Online]. Available: https://arxiv.org/abs/2503.11646

[19] Y. Wu, X. Chen, Y. Chen, H. Sadeghian, F. Wu, Z. Bing, S. Haddadin, A. König, and A. Knoll, "Sharedassembly: A data collection approach via shared tele-assembly," *arXiv preprint arXiv:2503.12287*, 2025. [Online]. Available: https://arxiv.org/abs/2503.12287

[20] M. F. Xu and B. Mutlu, "Exploring the use of robots for diary studies," *arXiv preprint arXiv:2501.04860*, 2025. [Online]. Available: https://arxiv.org/abs/2501.04860

[21] O. Balalau, C. Conceição, H. Galhardas, I. Manolescu, T. Merabti, J. You, and Y. Youssef, "Graph integration of structured, semistructured, and unstructured data for data journalism," *arXiv preprint arXiv:2007.12488*, 2020. [Online]. Available: https://arxiv.org/abs/2007.12488

[22] Y. Fan, J. Ohme, and C. Neuberger, "Digital turn without digital methods? mapping the journey of journalism studies," *Digital Journalism*, vol. 13, pp. 1–27, 2025, online first. [Online]. Available: https://doi.org/10.1080/21670811.2025.2480106

[23] Y. Yuan, Y. Song, Z. Luo, W. Sun, and K. Kitani, "Embodied intelligence: A synergy of morphology, action, perception and learning," *ACM Computing Surveys*, vol. 57, no. 7, pp. 186:1–186:36, 2025.

[24] S. Chen, P.-L. Guhur, M. Tapaswi, C. Schmid, and I. Laptev, "Think global, act local: Dual-scale graph transformer for vision-and-language navigation," in *European Conference on Computer Vision*, ser. Lecture Notes in Computer Science, 2022.

[25] M. Ahn, A. Brohan, Y. Chebotar, *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," in *Proceedings of Robotics: Science and Systems*, 2022.

[26] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, *et al.*, "RT-1: Robotics transformer for real-world control at scale," *arXiv preprint arXiv:2212.06817*, 2022. [Online]. Available: https://arxiv.org/abs/2212.06817

[27] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23 716–23 736, 2022.

[28] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *arXiv preprint arXiv:2304.08485*, 2023. [Online]. Available: https://arxiv.org/abs/2304.08485

[29] B. Zhang, K. Li, Z. Cheng, Z. Hu, Y. Yuan, G. Chen, S. Leng, Y. Jiang, H. Zhang, X. Li, P. Jin, W. Zhang, F. Wang, L. Bing, and D. Zhao, "VideoLLaMA 3: Frontier multimodal foundation models for image and video understanding," *arXiv preprint arXiv:2501.13106*, 2025. [Online]. Available: https://arxiv.org/abs/2501.13106

[30] F. Li, R. Zhang, H. Zhang, Y. Guo, W. Li, W. Zhu, T. Li, R. Jin, S. Li, X. Li, *et al.*, "LLaVA-NeXT-Interleave: Tackling multi-image, video, and 3D in large multimodal models," *arXiv preprint arXiv:2407.07895*, 2024. [Online]. Available: https://arxiv.org/abs/2407.07895

[31] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022, accessed: 2025-09-14. [Online]. Available: https://arxiv.org/abs/2212.04356

[32] P. Herrero-Diz, D. Varona-Aramburu, and M. Pérez-Escolar, "Debunking news as a journalistic genre: From the inverted pyramid to a circular writing model," *International Journal of Communication*, vol. 18, pp. 1634–1656, 2024. [Online]. Available: https://ijoc.org/index.php/ijoc/article/view/21570

[33] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

[34] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019. [Online]. Available: https://arxiv.org/abs/1907.11692