Multi-Agent Video Prediction: Self-Correcting Conditional Frames for Dynamic Scene Forecasting

Qixin Zhang Ajay Kumar Gurumadaiah Zhi-Li Zhang
University of Minnesota
{zhan8548, gurum021, zhang089}@umn.edu

Abstract

Transmission latency significantly degrades user quality of experience in real-time interactive perception systems. In remote driving, maintaining reliable visual feedback is critical for safe operation under dynamic network variability. Although video prediction offers a promising approach to compensate for short-term transmission delays and approximate near-zero-latency streaming, prediction-only methods remain vulnerable in highly dynamic scenes, especially when newly emerged objects appear during uplink outages. To address these challenges, we propose a multi-agent video prediction framework that combines continuous edge-side video prediction with lightweight mask-guided conditional frame reconditioning. The framework consists of three role-specialized agents: a continuous prediction agent for low-latency visual continuity, a vehicle-side trigger agent for detecting newly appeared objects, and a conditional reconditioning agent that repairs the predictor conditioning state using sparse mask guidance. This design enables semantic recovery of exogenous scene changes without requiring full-frame retransmission. We validate the proposed framework through extensive experiments on benchmark video data under realistic 5G communication traces. Results show that our method improves semantic recovery of novel objects while preserving perceptual quality and practical runtime efficiency under network-induced disruptions.

1. Introduction

Real-time visual communication has become a core component of modern interactive systems, enabled by advances in wireless networking, cloud infrastructure, and edge computing [18]. Applications such as remote collaboration, interactive streaming, surveillance, and online gaming rely on continuous video transmission to support real-time perception and interaction, making low end-to-end latency critical for smooth and reliable operation. This requirement becomes even more important in safety-critical and human-in-the-loop scenarios, including telesurgery, remote robotic manipulation, and remote vehicle operation, where operators depend on timely visual feedback to make accurate control decisions [4, 29]. However, delays introduced by network transmission, video encoding, and distributed processing can degrade situational awareness and response time, making latency mitigation a key challenge in teleoperation systems [15, 27].

Recent advances in deep generative models, including diffusion-based and autoregressive architectures, have significantly improved the fidelity and temporal coherence of predicted video frames. Models such as MCVD [25] and ExtDM [30] demonstrate the ability to synthesize plausible future frames conditioned on past observations. However, their effectiveness in real-world teleoperation scenarios remains limited due to the highly dynamic and interactive nature of environments such as road traffic. Small prediction errors can accumulate over time, leading to drift, inconsistent object motion, or unrealistic scene evolution, particularly when the model lacks a deeper understanding of the underlying scene dynamics

To address these limitations, we propose a Multi-Agentic Video Prediction framework that introduces collaborative agents to improve prediction robustness in dynamic environments. Instead of relying on a single predictive model, our framework employs specialized agents that maintain visual continuity and restore missing scene semantics during communication gaps. A prediction agent first generates future frames, while auxiliary agents analyze scene dynamics and generate refined conditional frames to correct accumulated drift or noise. This collaborative mechanism enables the system to maintain consistency with the underlying scene structure and motion, leading to more reliable video prediction under real-world teleoperation conditions. Figure 1 provides a high-level overview of the proposed multi-agent framework.

Motivated by these challenges, we propose a Multi-Agentic Video Prediction framework that integrates continuous prediction, novel-object triggering, and conditional reconditioning to repair the predictor conditioning state.

This collaborative mechanism improves temporal consistency and robustness in dynamic environments. The key contributions of this work are summarized as follows:

- We propose a multi-agent reconditioning framework for predicted video streaming that combines continuous edge prediction with lightweight mask-guided inpainting to inject newly emerged objects into the conditioning state.
- We identify and formalize a structural failure case in prediction-only streaming under open-world communication gaps, termed Novel-Object Blindness (NOB), which motivates the need for reconditioning.
- We provide qualitative and quantitative evidence under realistic 5G uplink traces that the proposed multi-agent framework improves semantic recovery while maintaining perceptual quality and practical runtime overhead.

2. Related Work

2.1. Multi-Agent Coordination in Embodied Systems

Embodied AI research has established interactive environments and benchmark suites for grounding perception, planning, and action under partial observability, including Habitat, ALFRED, and TEACh [16, 19, 22]. Recent agentic frameworks further explore role specialization and inter-agent communication for complex tasks, including CAMEL, AutoGen, and Voyager [11, 26, 28]. These studies show that decomposing a task into specialized agent roles can improve adaptability and robustness in open-ended settings.

Most of these systems, however, optimize for task-level outcomes such as instruction completion, dialogue quality, or long-horizon exploration [11, 16, 26, 28]. Their coordination primitives are typically designed for planning and decision-making over symbolic or text-centric state representations, rather than for preserving semantic fidelity in a continuously generated visual stream [11, 28]. In communication-constrained teleoperation, the central requirement is different: agents must coordinate online to maintain both temporal continuity and object-level scene correctness under intermittent observations [5, 10].

However, existing multi-agent literature is primarily centered on task completion, dialogue planning, or long-horizon policy execution. It does not directly address semantic reliability in predictive video streams under communication outages. Our work targets this gap by framing communication-gap recovery as role-specialized coordination between a continuity agent and a state-correction agent.

2.2. Video Prediction Under Partial Observability

Autoregressive video prediction methods extrapolate future frames from a fixed history window. Early recurrent models such as ConvLSTM and PredNet established temporal latent-state propagation [13, 21], and subsequent work improved motion-content modeling for longer horizons [24]. Diffusion-based video models, including Video Diffusion Models and MCVD, have further improved perceptual quality in conditional generation [7, 25].

Frame interpolation methods such as Super-SloMo and RIFE produce high-quality in-between frames when both temporal anchors are available [8, 9]; however, they are not applicable during real-time transmission outages where future observations are unavailable. More broadly, high-fidelity diffusion models [2, 6, 17] prioritize perceptual realism but do not explicitly guarantee object-level semantic correctness under exogenous scene changes.

Under communication gaps, this distinction is critical. A method can produce visually plausible and temporally smooth predictions while still failing to update scene state when novel entities emerge outside the conditioning history. This failure is structural rather than cosmetic: the issue is not merely reconstruction error, but the absence of an explicit mechanism for exogenous evidence injection into the predictive state. As a result, standard quality metrics alone may understate safety-relevant semantic omissions in teleoperation contexts [5, 10].

Overall, prior video modeling methods emphasize realism or temporal smoothness under closed-history conditioning. They do not provide an explicit mechanism for injecting newly observed exogenous evidence into an ongoing predictive stream during communication gaps.

2.3. Communication-Constrained Teleoperation and Streaming

Prior teleoperation and streaming research has extensively studied low-latency transmission and network robustness, including DASH-style adaptation, edge-assisted delivery, and delay/packet-loss analysis [1, 5, 10, 14, 20, 23]. These methods are essential for system-level reliability, but they largely treat visual content as passively transmitted data.

In contrast, our focus is semantic reliability under communication constraints: when outages occur, the stream must remain both temporally continuous and semantically correct. This requires coordination between predictive generation and targeted state correction, rather than transmission adaptation alone.

3. Problem Statement

Video prediction plays a critical role in latency-sensitive teleoperation systems where future frames must be estimated when new visual observations are delayed or unavailable. However, existing video prediction methods are primarily designed for offline forecasting settings and do not explicitly address the challenges introduced by communication latency, dynamic environments, and incomplete observations. To better understand these limitations, we for-

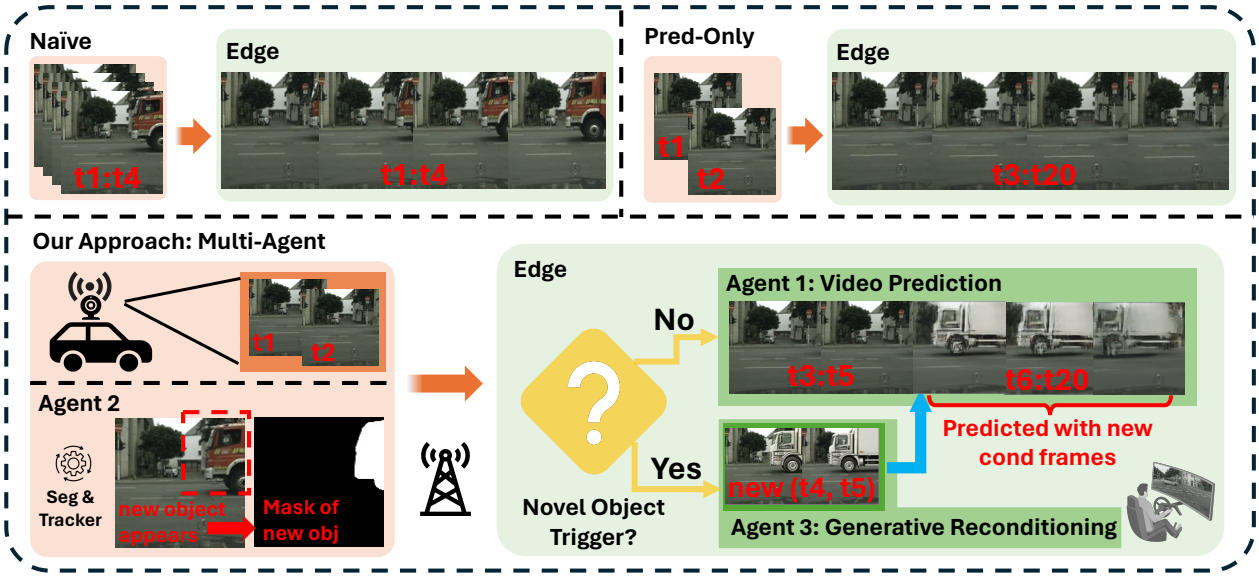

Figure 1. Architecture of Our Multi-Agent Approach.

mulate the following key research questions that guide the design of our approach.

Q1. Edge Extrapolation and the Single-Agent Limitation: In latency-compensated teleoperation systems, video prediction often relies on the most recently received frame when network delays prevent timely updates. Existing methods typically use a single predictive agent that extrapolates scene dynamics from this conditional frame using learned motion representations such as optical flow or latent motion features. While effective for short-term prediction, this approach becomes unreliable in dynamic environments. As predictions recursively depend on previously generated frames, errors accumulate over time, leading to temporal drift and inconsistent scene evolution.

Q2. Novel-Object Blindness under Communication Gaps: Another challenge arises during communication gaps when the teleoperation system fails to receive new frames from the remote environment. During these interruptions, prediction models generate frames based only on previously observed data, making them unaware of new objects or events that may appear in the scene. For example, in remote driving scenarios, new vehicles or pedestrians may enter the scene while communication is disrupted. As a result, the predicted frames may remain visually plausible but fail to reflect the true evolving state of the environment. We term this failure mode *Novel-Object Blindness* (NOB), where prediction-only rollout fails to recover newly appeared objects that are absent from the current conditioning history.

Q3. Uplink Instability and Perceptual Discontinuity in Teleoperation: In real-world teleoperation systems, video streams are transmitted over wireless networks such as 5G, where uplink bandwidth and latency can fluctuate due to channel variability and network congestion, as shown in Figure 2. These fluctuations can cause irregular frame delivery or temporary interruptions. During such disruptions, prediction models must generate frames without synchronized updates from the remote environment. When delayed frames arrive, discrepancies between predicted and actual scenes can produce abrupt visual changes, leading to perceptual discontinuities that degrade operator situational awareness.

Motivated by the above research questions, we define the problem of robust video prediction under communication latency and dynamic multi-agent environments. Let $X_{t-k:t} = \{X_{t-k}, \ldots, X_t\}$ denote a sequence of k+1 observed frames up to time t. The objective is to learn a prediction model that estimates future frames over a prediction horizon T that estimates the future frame sequence.

$$\hat{X}_{t+1:t+T} = f(X_{t-k:t}) \tag{1}$$

where T denotes the prediction horizon and $\hat{X}_{t+1:t+T}$ represents the predicted future frames. Recent approaches, including autoregressive and diffusion-based models, generate future frames by conditioning on previously observed frames and iteratively predicting subsequent frames.

$$\hat{X}_{t+i} = f(X_{t-k:t}, \hat{X}_{t+1:t+i-1}), \quad i = 1, \ldots, T \tag{2}$$

While these models have demonstrated promising results for short-term prediction in relatively structured environments, their performance degrades in highly dynamic scenarios.

In real-world applications such as remote driving and robotic teleoperation, environments often contain multiple interacting agents (e.g., vehicles, pedestrians, cyclists) whose motions evolve unpredictably. When prediction

models rely solely on previously predicted frames as conditioning inputs, small inaccuracies in motion or appearance can accumulate over time, leading to temporal drift, motion inconsistencies, and unrealistic scene evolution. This problem becomes more severe when prediction is used to compensate for communication latency, where predictions must remain stable and physically plausible for extended horizons.

$$\epsilon_{t+i} = X_{t+i} - \hat{X}_{t+i} \tag{3}$$

Consider a teleoperation video stream with intermittent uplink frame delivery. Let X_t denote the ground-truth frame at time t, and let c_t denote the conditioning state available at the edge for future prediction. In a prediction-only system, when new uplink observations are missing, the edge predictor rolls out future frames autoregressively from stale conditioning inputs. Formally, the predicted stream is generated as

$$\hat{X}_{t+i} = f_{\text{pred}}(c_t, \hat{X}_{t+1:t+i-1}), \quad i = 1, \ldots, T. \tag{4}$$

where f_{pred} denotes the video prediction model and T is the rollout horizon.

This formulation is effective for preserving short-term temporal continuity, but it assumes that all semantically relevant scene content is already represented in the conditioning state. In practice, this assumption fails in open-world teleoperation. If a previously unseen object enters the scene during a communication gap, that object is absent from the edge-side conditioning history and therefore cannot be inferred reliably by prediction alone. As a result, the predicted stream may remain temporally smooth while becoming semantically incomplete.

Our goal is therefore not only to maintain visual continuity during uplink outages, but also to restore missing scene semantics when such exogenous events occur. To do so, we seek a mechanism that (i) preserves the low-latency benefits of edge-side prediction, (ii) avoids the bandwidth cost of retransmitting full frames, and (iii) updates the predictor using sparse corrective evidence when a novel object appears. In our framework, this corrective evidence is provided as a lightweight object mask and is used to repair the predictor conditioning state before autoregressive rollout continues.

4. Agentic Video Prediction Framework

4.1. System Overview

We propose an agentic video prediction framework for maintaining continuous visual feedback under intermittent uplink transmission while correcting semantic omissions caused by prediction-only rollout. The framework consists of three role-specialized agents deployed across the vehicle and the edge: (i) a Continuous Prediction Agent (Agent 1) on the edge, (ii) a Novel-Object Trigger Agent (Agent 2) on the vehicle, and (iii) a Conditional Frame Reconditioning Agent (Agent 3) on the edge. Figure 3 illustrates the overall pipeline.

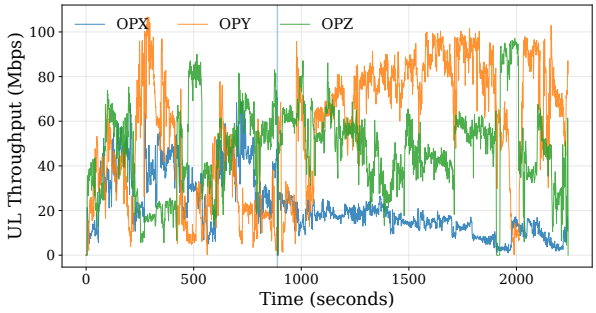

Figure 2. Uplink throughput over time in a 5G network for three operators in the U.S.

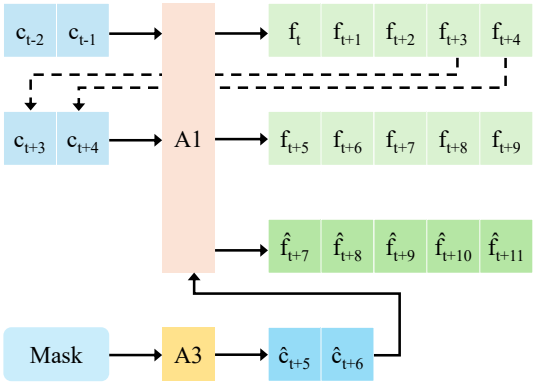

Figure 3. Multi-agent prediction and reconditioning pipeline on the edge. Notation: c_t conditioning frames; $\hat{c}_t$ repaired conditioning frames; f_t future frames predicted by $A1$; $\hat{f}_t$ reconditioned future frames.

Agent 1 is responsible for temporal continuity. In our implementation, it is instantiated with MCVD, although the framework itself is predictor-agnostic. Agent 2 monitors the live onboard stream and detects newly appeared objects using segmentation and tracking; in our experiments, we use a YOLO-based instance segmentation and tracking pipeline. Agent 3 performs mask-guided conditional generation to repair the predictor conditioning state, implemented with MagicQuil [12] in our system.

The framework follows an event-driven control policy. When fresh uplink frames are available, the edge updates the conditioning state using the received observations and performs prediction only when necessary to bridge short gaps. When expected frames are missing, Agent 1 rolls out future frames autoregressively to preserve low-latency visual continuity. If a novel object appears during such a missing-frame interval, Agent 2 transmits a compact mask

event to the edge, and Agent 3 uses this signal to repair the conditioning state used by Agent 1. Prediction then resumes from the repaired state. In this way, the three agents play complementary roles: Agent 1 preserves continuity, Agent 2 supplies sparse semantic evidence, and Agent 3 injects that evidence back into the predictor state without requiring full-frame retransmission.

4.2. Prediction Mechanism

Let $c_t = \{X_{t-1}, X_t\}$ denote the conditioning state maintained at the edge at prediction time t. More generally, the framework assumes a fixed-length conditioning context required by the underlying predictor; in our implementation with MCVD, this context consists of two frames. Given c_t, the Continuous Prediction Agent A_1 generates a horizon-T future segment autoregressively:

$$\hat{X}_{t+i} = f_{\text{MCVD}}(c_t, \hat{X}_{t+1:t+i-1}), \quad i = 1, \ldots, T. \tag{5}$$

This horizon-based formulation matches the inference interface of the predictor, while the generated segment is treated as a provisional continuation that can be replaced whenever fresher observations arrive.

The role of A_1 is to maintain a continuously available visual stream under incomplete frame delivery. Whenever new uplink frames are received, the edge refreshes the conditioning state using the latest ground-truth observations. When the incoming burst is incomplete, A_1 immediately predicts the missing suffix from the most recent valid conditioning state. Thus, the predictor is not used as a standalone forecasting module, but as an online continuity mechanism whose state is repeatedly updated by the frame-arrival process.

We formalize this behavior as a sliding-window state transition. When a new ground-truth frame arrives, it replaces the oldest element in the conditioning state. When no new observation is available, the predictor advances using its own generated outputs. For the two-frame case used in our experiments, the update rule is

$$c_{t+1} = \begin{cases} \{X_t, X_{t+1}\}, & \text{if ground-truth } X_{t+1} \text{ arrives,} \\ \{\hat{X}_t, \hat{X}_{t+1}\}, & \text{otherwise.} \end{cases} \tag{6}$$

This recursive update is identical to the prediction-only baseline during ordinary outages. Our contribution is not to modify the predictor itself, but to intervene on its conditioning state when new semantic evidence becomes available, so that subsequent rollouts remain both temporally smooth and semantically updated.

4.3. Novel-Object Trigger Agent

The Novel-Object Trigger Agent A_2 runs on the vehicle and monitors the live camera stream for newly appeared instances. We use a YOLO-based instance segmentation model with tracking to maintain object identities across frames. A novel object is defined as an instance whose track ID is absent from the recent history window. When such an event is detected at time t^*, A_2 sends a compact binary mask M_{t^*} (and its timestamp) to the edge. This event-driven signaling is bandwidth-lightweight compared with transmitting full frames and is robust to uplink constraints.

Formally, let $\mathcal{I}_t$ be the set of tracked instance IDs at time t, and let $\mathcal{H}_t = \bigcup_{j=1}^{K} \mathcal{I}_{t-j}$ denote the ID history over the last K frames. A novel object is detected when $\mathcal{I}_t \setminus \mathcal{H}_t \neq \emptyset$. The trigger agent then generates a mask M_t by unioning the pixel regions of all newly appeared instances:

$$M_t = \bigcup_{o \in \mathcal{I}_t \setminus \mathcal{H}_t} \mathbb{1}_{\text{mask}}(o). \tag{7}$$

Only the binary mask and a timestamp are transmitted. This design keeps the uplink payload small (on the order of kilobytes) and makes triggering feasible even under severe bandwidth constraints.

If multiple novel objects appear within a short interval, their masks are merged into a single trigger payload. If no new object is detected, the agent remains silent, thereby avoiding unnecessary communication. This selective signaling is critical for preserving uplink capacity while still enabling semantic correction when it matters.

The trigger message is buffered at the edge and aligned to the closest predicted-frame timestamp. This alignment ensures that the reconditioning agent operates on a temporally consistent predictor state, reducing visual discontinuities after correction.

4.4. Conditional Frame Generation

The Conditional Frame Reconditioning Agent A_3 is activated only when a novel-object trigger is received during a prediction interval. Its purpose is not to replace the displayed video stream directly, but to repair the internal conditioning state of the predictor so that the missing object can be propagated by subsequent rollouts. Let t^* denote the trigger time, and let τ be the closest predicted timestamp available at the edge. In our two-frame instantiation, the reconditioning target is the aligned conditioning pair $\{\hat{X}_{\tau-1}, \hat{X}_\tau\}$ together with the transmitted mask M_{t^*}.

Using MagicQuill, A_3 performs mask-guided conditional generation on the aligned predicted frames to obtain repaired conditioning frames:

$$\tilde{X}_{\tau-1} = f_{\text{MQ}}(\hat{X}_{\tau-1}, M_{t^*}), \qquad \tilde{X}_\tau = f_{\text{MQ}}(\hat{X}_\tau, M_{t^*}). \tag{8}$$

More generally, the number of frames to be reconditioned follows the conditioning length required by Agent 1. In our implementation, because MCVD uses a two-frame context, we repair two consecutive predicted frames around the

aligned trigger time. This keeps the repaired state compatible with the predictor interface and preserves short-range temporal coherence before autoregressive rollout resumes.

The repaired conditioning state is then used to reinitialize the predictor:

$$\tilde{c}_\tau = \{\tilde{X}_{\tau-1}, \tilde{X}_\tau\}. \tag{9}$$

Agent 1 subsequently continues prediction from the repaired state as

$$\hat{X}_{\tau+i} = f_{\text{MCVD}}(\tilde{c}_\tau, \hat{X}_{\tau+1:\tau+i-1}), \quad i = 1, \ldots, T. \tag{10}$$

A key design choice is that the repaired frames are injected into the predictor state rather than directly substituted into the output stream. Direct display replacement would create a mismatch between the visible frame and the latent temporal state from which future predictions are generated, often causing abrupt drift immediately after correction. By instead repairing the conditioning state, the framework allows A_1 to absorb the newly introduced object and propagate it forward in a temporally coherent manner. As a result, the additional system overhead is limited to sparse mask transmission and occasional reconditioning inference, while preserving the bandwidth advantages of prediction-based teleoperation.

5. Evaluation and Discussion

5.1. Experimental Setup

Dataset and Evaluation Scenarios We evaluate our approach on video sequences from the Cityscapes dataset [3], which provides real-world urban driving scenes with naturally occurring dynamic objects. The videos include spontaneous object appearances (e.g., pedestrians or vehicles entering the field of view), enabling evaluation under open-world conditions without synthetic insertion. We evaluate multiple sequences containing diverse novel-object events to ensure result robustness.

For each sequence, the first two frames are used as conditioning frames to initialize the predictive model. Subsequent frames evolve according to the original video dynamics. To ensure fair comparison across methods, frame counts and conditioning windows are controlled consistently when computing reconstruction metrics.

Communication Simulation. To emulate realistic teleoperation conditions, we replay uplink throughput traces collected from real-world 5G deployments, as shown in Figure 2. The trace dataset consists of approximately 18 hours of uplink measurements recorded at 100 ms intervals under diverse mobility conditions, including handover events and bandwidth fluctuations.

During simulated outage periods (uplink $\leq$ 5 Mbps), full-frame transmission is constrained by measured instantaneous bandwidth, and the edge must rely on predictive streaming to maintain visual continuity.

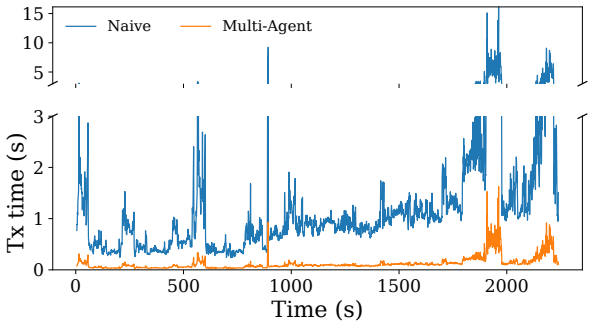

Figure 4. Instantaneous transmission time under the OPX uplink trace for the Naive approach and our proposed method.

Novel Object Events and Metrics. Novel-object events are identified using a vehicle-side YOLO-based segmentation and tracking module. When a previously unseen object instance enters the scene, a detection trigger is generated and transmitted to the edge as a lightweight mask signal. For evaluation, the timestamp of the first ground-truth appearance of the object in the video is recorded as the reference time.

To quantify Novel-Object Blindness, we define Scene Recovery Time (SRT) as the interval between (i) the ground-truth first appearance of a novel object and (ii) the first frame in the multi-agent output stream where the object becomes visible. SRT is measured within each 20-frame segment and directly captures semantic recovery behavior during communication gaps. In addition to SRT, we report standard video quality metrics, including FVD, LPIPS, MSE, PSNR, and SSIM, to assess overall perceptual fidelity.

All experiments are conducted on a single NVIDIA RTX A6000 GPU to emulate edge-side inference.

5.2. Semantic Recovery through Conditional Frame Generation

We evaluate how the proposed multi-agent framework restores semantic consistency when novel objects emerge during uplink interruptions. Figure 5 provides a qualitative comparison among ground truth, prediction-only streaming, and our multi-agent framework. The first row shows the ground-truth sequence, the second row corresponds to prediction-only streaming (Agent 1), and the third row shows our multi-agent framework (prediction + mask-guided reconditioning).

The predictive model is initialized with the first two frames (t=1 and t=2) as conditioning inputs. Upon receiving these frames, the edge predictor autoregressively generates future frames (t=3–12) in the absence of additional observations. At t=9, a previously unseen white vehicle enters the scene from the lateral direction while the ego vehicle con-

tinues moving forward. In the ground-truth sequence, this crossing vehicle introduces a potential collision risk if the ego vehicle does not decelerate.

However, in the prediction-only stream, the white vehicle is absent despite continued temporal smoothness of the generated frames. Since the object was not present in the initial conditioning window, the autoregressive predictor fails to incorporate it into the evolving scene, resulting in a semantically incomplete video stream. In teleoperation, an operator observing such predicted frames would not be aware of the crossing vehicle and may fail to adjust speed accordingly.

In contrast, our multi-agent framework receives a mask of the newly detected white vehicle from the vehicle-side perception module. A lightweight inpainting agent generates an updated conditioning frame that incorporates the new object, after which the predictive agent is reconditioned and resumes autoregressive generation. As shown in the third row of Figure 5, the crossing vehicle is correctly introduced into subsequent frames, aligning the predicted stream with scene evolution.

To quantify this effect, Table 1 reports Scene Recovery Time (SRT) statistics under novel-object appearance. SRT measures the interval between the ground-truth first appearance of a novel object and its first visible reconstruction in the output stream. Prediction-only streaming exhibits a mean SRT of 500 ms, reflecting prolonged semantic blindness during communication gaps. Our multi-agent framework reduces the mean SRT to 275 ms, with substantially lower minimum recovery latency. These results confirm that the proposed reconditioning mechanism effectively shortens NOB duration.

We further evaluate overall perceptual fidelity in Table 2. The proposed approach achieves a substantial FVD improvement, indicating closer alignment with ground-truth video dynamics at the sequence level. Pixel-level metrics such as PSNR, SSIM, and MSE show only marginal differences relative to prediction-only streaming. These small variations arise from corrective object insertion during reconditioning but do not materially degrade visual continuity. Overall, the results show that terminating Novel-Object Blindness improves semantic consistency without significantly compromising low-level reconstruction quality.

Together, the qualitative and quantitative results demonstrate that the proposed multi-agent reconditioning framework restores semantic consistency more effectively than prediction-only streaming in open-world settings, with minimal impact on overall perceptual fidelity. We note that SRT focuses on semantic recovery within 20-frame segments and does not directly represent end-to-end operator response latency.

Table 1. Scene Recovery Time (SRT) statistics under novel object appearance (FPS=20), measured within each 20-frame segment in the output stream.

Method	Mean SRT (ms)	Min–Max (ms)
Pred-only	500	500–900
Multi-Agent	**275**	50–500

Table 2. Quantitative comparison on video prediction. Δ denotes (Ours - Baseline).

Metric	Pred-only	Ours	Δ
FVD $\downarrow$	437.56	178.75	-258.81
LPIPS $\downarrow$	0.0579	0.0724	0.0145
MSE $\downarrow$	0.00272	0.00347	0.00075
PSNR $\uparrow$	25.69	24.63	-1.06
SSIM $\uparrow$	0.9207	0.8936	-0.0270

5.3. Runtime and System Overhead

We finally analyze the runtime overhead of the proposed framework under communication gaps. Table 3 presents a latency breakdown for recovering a 20-frame outage under a 5 Mbps uplink constraint.

Under the naïve streaming strategy, all 20 frames must be transmitted directly, resulting in a total latency of 3280 ms. Prediction-only streaming reduces this to 1228 ms by transmitting only two conditioning frames and generating the remaining frames at the edge. Our multi-agent framework incurs a total latency of 1441 ms, which includes additional costs for mask generation, mask transmission, and lightweight inpainting.

Importantly, the added overhead relative to prediction-only streaming is modest while remaining substantially lower than the naïve full-frame transmission baseline. The additional latency corresponds to the corrective reconditioning step that enables termination of Novel-Object Blindness. Given the significant Scene Recovery Time reduction demonstrated in Section 5.2, this overhead represents a favorable trade-off between semantic reliability and transmission efficiency.

5.4. Communication-Constrained Deployment Analysis

We also analyze whether the proposed multi-agent pipeline remains deployable under realistic uplink perturbations. Figure 4 compares the instantaneous transmission time of naïve full-frame streaming and our multi-agent framework under the OPX uplink trace. Under the naïve strategy, each frame is transmitted directly over the uplink. As bandwidth fluctuates, transmission time increases sharply in low-throughput intervals, producing pronounced latency spikes that correspond to outage periods.

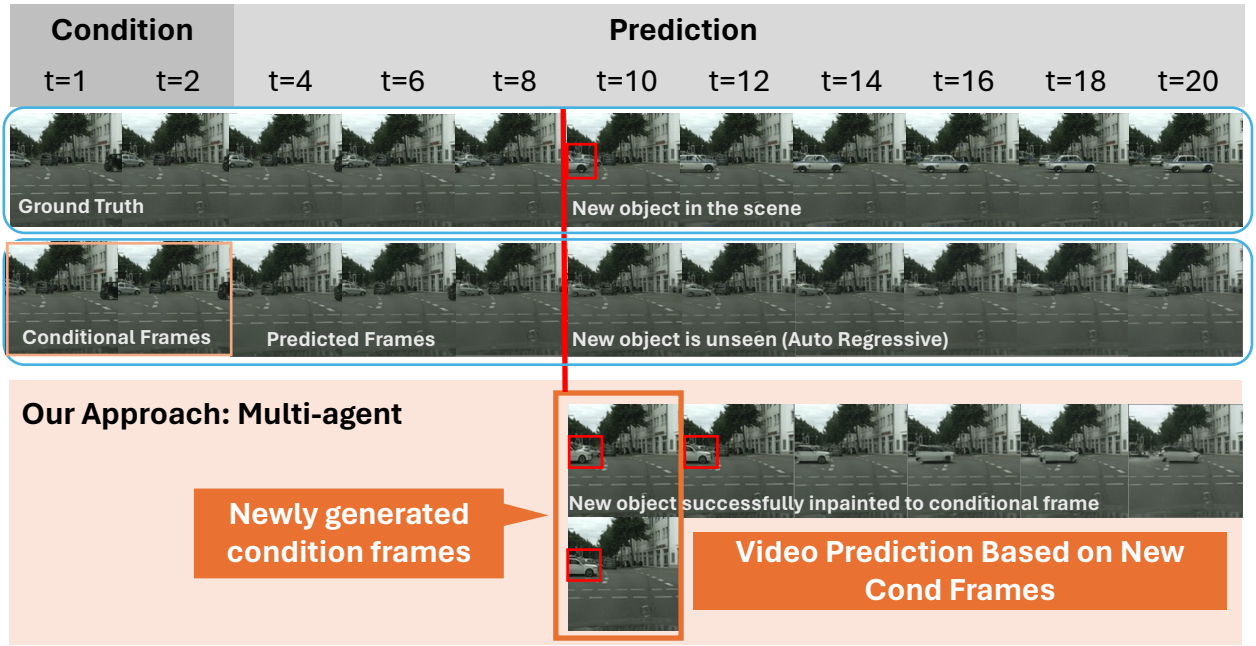

Figure 5. Qualitative comparison among ground truth, prediction-only streaming, and our approach.

Table 3. Latency breakdown for recovering a 20-frame gap (uplink=5 Mbps, image=100 KB/frame, mask=2 KB).

Method	Comp.	Time (ms)	Total (ms)
Naive	Transmit 20 frames	3280	**3280**
Pred-only	Transmit 2 cond frames	328	**1228**
	Predict 18 frames	900	
Multi-Agent	Transmit 2 cond frames	328	**1441**
	Generate mask	20	
	Transmit mask	3	
	Inpaint 2 frames	290	
	Predict 16 frames	800	

In contrast, our framework decouples stream continuity from instantaneous uplink bandwidth. During perturbations, only conditioning frames and lightweight mask signals are transmitted, while intermediate future frames are generated at the edge. As shown in Figure 4, this design substantially suppresses transmission-time bursts relative to naïve streaming.

These results indicate that the proposed multi-agent reconditioning pipeline can sustain low-latency operation under communication constraints, supporting practical edge deployment without continuous high-bandwidth connectivity.

Overall, the results indicate that the proposed multi-agent design achieves timely semantic recovery with marginal system overhead, making it suitable for real-time edge deployment under realistic 5G constraints.

6. Conclusion

This paper studies semantic reliability in edge-assisted predictive video streaming for teleoperation under realistic 5G uplink perturbations. We show that although autoregressive prediction preserves short-term visual continuity during communication gaps, it suffers from a structural failure mode under open-world scene evolution: newly appeared objects outside the conditioning history may be systematically omitted. We formalize this phenomenon as Novel-Object Blindness (NOB) and address it with a multi-agent reconditioning framework that couples continuous edge prediction with mask-guided lightweight inpainting.

Experiments under real 5G uplink traces show that the proposed design improves semantic recovery under communication gaps while preserving overall perceptual quality and real-time feasibility. In particular, Scene Recovery Time is substantially reduced compared with prediction-only streaming, and system overhead remains modest relative to naive full-frame transmission. These results indicate that semantic consistency can be improved without sacrificing practical deployability in edge teleoperation settings.

Acknowledgments

The research was supported in part by NSF awards CNS-2321531, DMS-2436333 and ITE-2453815 and MnDOT/LRRB grant.

References

[1] Abdelhak Bentaleb, Bayan Taani, Ali C Begen, Christian Timmerer, and Roger Zimmermann. A survey on bitrate

adaptation schemes for streaming media over http. *IEEE Communications Surveys & Tutorials*, 21(1):562–585, 2018. 2

[2] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22563–22575, 2023. 2

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 6

[4] Tomoya Hatano, Tatsuya Shimada, and Tomoaki Yoshida. Evaluation of end-to-end delay requirement for remote, precise pick-up action of miniature crane. *IEICE Communications Express*, 12(5):242–248, 2023. 1

[5] Sandra Hirche and Martin Buss. Packet loss effects in passive telepresence systems. In *2004 43rd ieee conference on decision and control (cdc)(ieee cat. no. 04ch37601)*, pages 4010–4015. IEEE, 2004. 2

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[7] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022. 2

[8] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *European conference on computer vision*, pages 624–642. Springer, 2022. 2

[9] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9000–9008, 2018. 2

[10] Burak Kizilkaya, Olaoluwa Popoola, Guodong Zhao, and Muhammad Ali Imran. 5g-based low-latency teleoperation: Two-way timeout approach. In *Annual Conference Towards Autonomous Robotic Systems*, pages 470–481. Springer, 2023. 2

[11] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for "mind" exploration of large language model society. *arXiv preprint arXiv:2303.17760*, 2023. 2

[12] Zichen Liu, Yue Yu, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Wen Wang, Zhiheng Liu, Qifeng Chen, and Yujun Shen. Magicquill: An intelligent interactive image editing system. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13072–13082, 2025. 4

[13] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016. 2

[14] Pavel Mach and Zdenek Becvar. Mobile edge computing: A survey on architecture and computation offloading. *IEEE communications surveys & tutorials*, 19(3):1628–1656, 2017. 2

[15] Mihir Mody, Pramod Swami, and Pavan Shastry. Ultra-low latency video codec for video conferencing. In *2014 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, pages 1–5. IEEE, 2014. 1

[16] Aishwarya Padmakumar, Jesse Thomason, Michael Murray, Parisa Kordjamshidi, Raymond Mooney, Peter Stone, and Prithviraj Ammanabrolu. Teach: Task-driven embodied agents that chat. *arXiv preprint arXiv:2110.00534*, 2021. 2

[17] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2

[18] Toshio Sato, Yutaka Katsuyama, Zheng Wen, Xin Qi, Kazuhiko Tamesue, Wataru Kameyama, Yuichi Nakamura, Jiro Katto, and Takuro Sato. Compensation of communication latency using video prediction in remote monitoring systems. *IEICE Proceedings Series*, 79(P2-32), 2023. 1

[19] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9339–9347, 2019. 2

[20] Weisong Shi, Jie Cao, Quan Zhang, Youhuizi Li, and Lanyu Xu. Edge computing: Vision and challenges. *IEEE internet of things journal*, 3(5):637–646, 2016. 2

[21] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015. 2

[22] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10740–10749, 2020. 2

[23] Thomas Stockhammer. Dynamic adaptive streaming over http– standards and design principles. In *Proceedings of the second annual ACM conference on Multimedia systems*, pages 133–144, 2011. 2

[24] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. In *International Conference on Learning Representations (ICLR)*, 2017. 2

[25] Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *Advances in neural information processing systems*, 35:23371–23385, 2022. 1, 2

[26] Guanzhi Wang, Yide Xie, Yunfan Jiang, Ajay Mandlekar, Chuchu Fan, Yuke Zhu, Anima Anandkumar, and Ruohan

Wang. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023. 2

[27] Hengchao Wang, Xu Zhang, Hao Chen, Yiling Xu, and Zhan Ma. Inferring end-to-end latency in live videos. *IEEE Transactions on Broadcasting*, 68(2):517–529, 2021. 1

[28] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Chi Liu, Ahmed Hassan Awadallah, Ryen White, and Doug Burger. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023. 2

[29] Yang Yu and Sanghwan Lee. Remote driving control with real-time video streaming over wireless networks: Design and evaluation. *IEEE Access*, 10:64920–64932, 2022. 1

[30] Zhicheng Zhang, Junyao Hu, Wentao Cheng, Danda Paudel, and Jufeng Yang. Extdm: Distribution extrapolation diffusion model for video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19310–19320, 2024. 1