IVA: Proactive Bitrate Orchestration for Multiparty Video Conferencing via Conversational Intent

Cheng Pan
The University of Hong Kong
Hong Kong, China
cpanpan@connect.hku.hk

Yuying Li
Simon Fraser University
British Columbia, Canada
yla924@sfu.ca

Cong Zhang
Shenzhen MSU-BIT University
Shenzhen, China
zhangcong@smbu.edu.cn

Edith C. H. Ngai*
The University of Hong Kong
Hong Kong, China
chngai@eee.hku.hk

Jiangchuan Liu
Simon Fraser University
British Columbia, Canada
jcliu@sfu.ca

Bo Li
Hong Kong University of Science and Technology
Hong Kong, China
bli@ust.hk

Baochun Li
University of Toronto
Ontario, Canada
bli@ece.toronto.edu

Abstract

Multimodal large language models (MLLMs) are increasingly integrated into video conferencing, but mostly for speech-centric tasks such as transcription and summarization. Conferencing adaptation remains dominated by signal-driven congestion control that reacts to bandwidth changes without understanding *why* certain moments or streams will soon become quality-critical. This separation wastes predictive structure in conversation: text often reveals impending role shifts and interaction-mode transitions—for example, taking the floor or initiating screen sharing—seconds before the corresponding media and bandwidth demands materialize.

We present **I**ntent-aware **V**ideo conferencing **A**daptation (IVA), a cross-modal control framework that converts text-derived intent cues into proactive bitrate orchestration. IVA extracts *activity*, *control*, and *intent* semantics from live streams, maps them to per-stream priority weights via a gated–bias model, and solves constrained bitrate allocation to maximize semantic QoE under uplink/downlink and latency limits. IVA learns this semantics-to-priority mapping with reinforcement learning over trace-driven simulations. Experiments with real-world network traces and emulated conferencing scenarios show that IVA improves QoE over signal-only ABR baselines and a semantic-aware conferencing baseline, while preserving low end-to-end latency and stabilizing quality around intent-triggered transitions.

*Edith C. H. Ngai is the corresponding author.

NOSSDAV '26, Hong Kong, Hong Kong

ACM ISBN 979-8-4007-2534-0/2026/04
https://doi.org/10.1145/3798065.3798068

CCS Concepts

• **Information systems** → **Multimedia streaming**.

Keywords

Intent-aware Adaptation, Multiparty Video Conferencing, Proactive Bitrate Orchestration

ACM Reference Format:
Cheng Pan, Yuying Li, Cong Zhang, Edith C. H. Ngai, Jiangchuan Liu, Bo Li, and Baochun Li. 2026. IVA: Proactive Bitrate Orchestration for Multiparty Video Conferencing via Conversational Intent. In *The 36th edition of the Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV '26), April 04–08, 2026, Hong Kong, Hong Kong*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3798065.3798068

1 Introduction

Real-time video conferencing has become core infrastructure for remote work and collaboration [1], and recent measurement reports attribute a substantial fraction of Internet traffic to video services [2]. In parallel, multimodal large language models (MLLMs) have enabled a new generation of AI assistants embedded in conferencing platforms [3–6]. Commercial tools such as Fireflies.ai [7], Zoom AI [8], and Otter AI [9] provide live transcription, summarization, search, and translation. Despite their success, these assistants typically operate *outside* the media control loop: they help users understand or document a meeting, but they rarely influence how the system allocates bitrate and latency budget across streams in real time.

At the same time, modern conferencing quality is still largely determined by signal-driven adaptation. Platforms

built atop WebRTC [10] adjust media rates based on network measurements such as throughput, RTT, loss, and jitter. State-of-the-art adaptive bitrate (ABR) schemes [11, 12] refine this process by better estimating available bandwidth and coordinating rates across multiple streams. However, these methods remain fundamentally *reactive*: they respond after congestion emerges and typically treat streams as interchangeable units whose importance is inferred only indirectly (e.g., from recent packet statistics). This mismatch is most visible during semantically critical moments—for example, when a presenter begins screen sharing, when the meeting transitions from presentation to discussion, or when multiple participants unmute during Q&A. These are precisely the moments where users are least tolerant of quality degradation, yet conventional controllers have no mechanism to anticipate them.

A key observation motivates this paper: conversational text is not merely a transcript of *what was said*; it is also a real-time signal of *what will happen next*. Meeting dialogue frequently contains explicit or implicit intent cues that foreshadow role changes and interaction-mode transitions: "I'll share my slides now," "Any questions?," or "You can go ahead." Such cues often precede the corresponding system-level events by seconds, because users must still perform UI actions (e.g., clicking to share a window, locating a control, or unmuting), and other participants need time to react. This gap creates an opportunity: if the system can recognize intent early, it can proactively reshape bitrate allocation *before* bandwidth contention and quality collapse occur.

In this paper, we argue that conferencing adaptation should incorporate intent as a first-class control signal. We propose **I**ntent-aware **V**ideo conferencing **A**daptation (IVA), a cross-modal framework that bridges language understanding and bitrate orchestration. IVA runs alongside the conventional conferencing pipeline. It uses ASR and an MLLM to interpret live text streams and to infer three categories of semantics: (1) *activity* signals that describe what users are doing now (e.g., speaking activity and speech energy), (2) *control* signals that capture explicit UI actions (e.g., mute/unmute and attention/focus), and (3) *intent* signals that predict imminent transitions (e.g., taking the floor or initiating screen sharing). IVA converts these semantics into per-stream priority weights using a gated–bias model that treats control actions as structural operators and preserves intent contributions even when a participant is currently silent. Given these priorities, IVA solves a constrained optimization to allocate bitrates across sender–receiver pairs while respecting uplink and downlink capacity limits and latency requirements.

A central challenge is that the mapping from semantics to "who should be protected" is context-dependent. For example, in presentation mode, the system should protect the presenter and shared content, whereas in free discussion it should

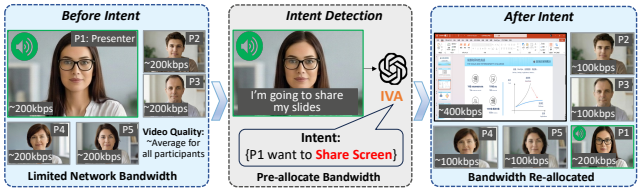

Figure 1: Bitrate Allocation Before vs. After Intent.

prioritize low latency and rapidly shift protection across short speaking turns. IVA addresses this by learning role–mode–specific priority parameters via reinforcement learning over trace-driven simulations: each role–mode group is trained to maximize a semantic QoE objective that rewards high quality for semantically important streams and penalizes instability, subject to bandwidth and latency constraints. This learning-based design allows IVA to adapt priority tradeoffs to different interaction structures without hard-coding per-scenario rules. Our original contributions in this paper are as follows.

- We identify a missing link in conferencing systems: conversational text contains predictive intent cues (e.g., role handoff, screen-sharing intent, and mode transitions) that precede quality-critical events, yet existing ABR controllers cannot exploit them.
- We design IVA, a cross-modal control framework that extracts activity, control, and intent semantics from live conferencing streams and uses a gated–bias priority model to drive constrained bitrate orchestration.
- We develop an RL-based method to learn role–mode–specific semantic priority parameters and evaluate IVA using trace-driven simulations and network emulation. Results show that intent-aware orchestration improves QoE over signal-driven baselines and a semantic-aware conferencing baseline, while maintaining low end-to-end latency around intent-triggered transitions.

2 Background & Motivation

Existing multiparty conferencing systems rely on signal-driven adaptation from network measurements such as throughput, RTT, and jitter. While effective at avoiding persistent congestion, these controllers remain fundamentally *reactive* and lack a principled mechanism to decide *which* streams to protect during semantically critical moments. Prior semantic communication work [13–16] reduces transmission overhead by prioritizing task-relevant semantics, but does not directly address conferencing QoE, where latency and temporal continuity are primary constraints. In meetings, however, semantics are often *predictive*: conversational text frequently reveals what will happen next.

Consider a five-participant meeting where a presenter announces an upcoming screen share (e.g., *"I'm going to*

Table 1: Intent-aware Events in Video Conferencing

Role-Mode	Event	Typical Intent Cue
Role Dynamics	Presenter speaking	"Let me explain this."
	Speaker handoff	"You can go ahead."
	Q&A interaction	"Any questions?"
	Audience interruption	"Sorry—quick question."
Mode Transitions	Start screen sharing	"I'll share my slides."
	Presentation mode	"Now I'll present."
	Discussion mode	"Let's discuss this."
	End presentation → Q&A	"That's all from me."

share my slides now."). Although intent is clear immediately, a signal-only controller does not react until sharing starts and bandwidth demand increases. Initiating sharing requires UI actions (clicking, selecting a window, starting capture), which we observe to take roughly 3–10 seconds, creating a predictive window. When sharing begins, reactive control often cuts bitrates abruptly, degrading presenter audio or shared content exactly when quality matters most. As shown in Fig. 1, IVA detects intent early, prioritizes shared content, preserves presenter quality, and smoothly de-prioritizes non-critical streams.

Tab. 1 highlights that (1) quality-critical events are often semantically signaled before they occur, and (2) stream importance depends on role and interaction mode, not only bandwidth trends. These observations motivate IVA: treat conversational intent as a first-class control signal for proactive bitrate orchestration.

3 System Overview

IVA targets multiparty video conferencing and allocates resources based on *semantic stream importance*. The key idea is to infer *what will happen next* from conversational signals and adjust bitrate allocation *before* contention causes quality degradation. Fig. 2 shows the IVA control loop integrated alongside a conventional conferencing pipeline.

At runtime, IVA ingests multiparty audio streams, user-interface (UI) events, and encoded video/transport signals. Audio is processed by VAD and ASR to produce a streaming transcript, which is fed into an MLLM for semantic interpretation. The combined ASR+MLLM delay is under 2.2 seconds [17], shorter than the empirically observed delays of role and interaction-mode transitions in Section 2, enabling proactive control. IVA extracts three semantic categories: **activity** (e.g., speaking activity, speech energy, effective speaking time), **control** (e.g., mute/unmute, application focus), and **intent** (e.g., floor requests, screen-sharing initiation, and role–mode transitions inferred from text). These signals expose interaction dynamics invisible to signal-only ABR.

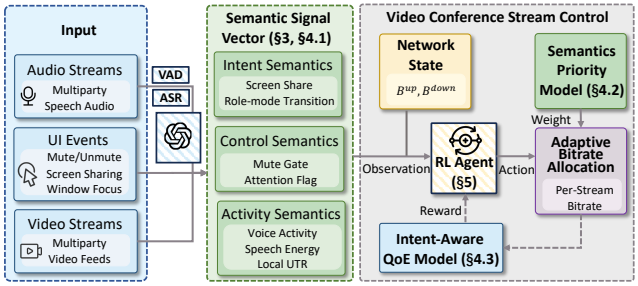

Figure 2: IVA Overview

IVA maps semantic signals to per-stream priority weights via a gated–bias model: control actions act as structural operators (e.g., mute as a hard gate), activity captures current salience, and intent biases priorities toward streams likely to become quality-critical even if currently silent. Given priorities, IVA solves constrained bitrate allocation under uplink/downlink capacity and latency limits. Because the semantics-to-priority mapping is context dependent, IVA learns role–mode–specific parameters via reinforcement learning over trace-driven simulations and selects the appropriate parameterization online as the inferred role–mode evolves.

4 Intent-Aware QoE Model

We model a multiparty conferencing session with n participants $\mathcal{N} = \{1, \ldots, n\}$ over discrete control intervals of length Δt seconds. At each $t \in \{1, \ldots, T\}$, each sender $i \in \mathcal{N}$ delivers a video stream to each receiver $j \in \mathcal{N} \setminus \{i\}$. Let $b_{ij}(t) \geq 0$ denote the bitrate allocated to stream $i \to j$ during t. Network conditions are summarized by the available uplink capacity $B_i^{\text{up}}(t)$ for sender i and the available downlink capacity $B_j^{\text{down}}(t)$ for receiver j (both in bitrate units). IVA operates by (1) extracting semantic features from multimodal conferencing signals, (2) mapping semantics to per-stream *priority scores*, and (3) allocating bitrates under bandwidth and latency constraints to maximize a *semantic QoE*.

4.1 Semantic Feature Definitions

At each interval t, IVA extracts three categories of semantic features for every directed stream $i \to j$.

(1) Activity semantics. Activity describes what sender i is doing *now* as inferred from recent audio/video. We define $\mathbf{x}_{ij}^{\text{act}}(t) = \left[v_i(t),\ E_i(t),\ U_i(t)\right]^\top$, where $v_i(t) \in \{0, 1\}$ is a voice-activity indicator (VAD) for sender i, $E_i(t) \in \mathbb{R}_+$ is a normalized speech-energy/loudness feature, and $U_i(t) \in [0, 1]$ is the sender's recent effective speaking time (e.g., local UTR) [18]. (We index by ij for notational uniformity, although these activity features are sender-side and do not depend on j.)

(2) Control semantics. Control captures explicit UI actions that affect media relevance or delivery. We use $m_i(t) \in \{0, 1\}$, $a_{ij}(t) \in \{0, 1\}$, where $m_i(t) = 1$ indicates sender

i is muted at time t (hard gating speech-driven importance), and $a_{ij}(t) = 1$ indicates receiver j is explicitly attending to sender i (e.g., pinned or spotlighted tile, active-speaker focus, or an explicit UI attention signal). If such fine-grained attention is unavailable, $a_{ij}(t)$ can be set to the receiver's application-focus flag as a coarse proxy.

(3) Intent semantics. Intent predicts what is likely to happen *next* over a short lookahead horizon (seconds), inferred from conversational text (ASR) and an MLLM. We define $\mathbf{x}_{ij}^{\text{int}}(t) = \left[p_i^{\text{share}}(t),\ p_i^{\text{floor}}(t),\ \boldsymbol{\pi}_i(t)\right]^\top$. Here $p_i^{\text{share}}(t) \in [0,1]$ is the probability that sender i will initiate screen sharing soon, and $p_i^{\text{floor}}(t) \in [0,1]$ is the probability that i will request/take the floor soon. In addition, $\boldsymbol{\pi}_i(t) \in \Delta^{G-1}$ is an MLLM-predicted posterior over G role–mode classes (e.g., *primary speaker in presentation*, *audience in discussion*), where Δ^{G-1} is the probability simplex.

4.2 Semantic Priority Model

IVA converts semantic features into per-stream priority scores that reflect *semantic protection importance*. Because the mapping depends on role and interaction mode, IVA maintains a role–mode–specific parameter set $\Theta^{(g)} = \{\boldsymbol{\theta}_{\text{act}}^{(g)}, \boldsymbol{\theta}_{\text{int}}^{(g)}, \kappa^{(g)}\}$ for each role–mode class $g \in \{1, \dots, G\}$.

Role–mode–conditioned score. For each class g, we compute an unnormalized priority score for stream $i \to j$ as:

$$w_{ij}^{(g)}(t) = \underbrace{(1 - m_i(t))\left(\boldsymbol{\theta}_{\text{act}}^{(g)}\right)^\top \mathbf{x}_{ij}^{\text{act}}(t)}_{\text{activity gated by mute}} + \underbrace{\text{ReLU}\left(\left(\boldsymbol{\theta}_{\text{int}}^{(g)}\right)^\top \mathbf{x}_{ij}^{\text{int}}(t)\right)}_{\text{intent bias (not muted)}} + \underbrace{\kappa^{(g)}\, a_{ij}(t)}_{\text{attention bias}}, \quad (1)$$

where $(1 - m_i(t))$ is a hard mute gate applied to activity (a muted sender should not be prioritized purely because of current speech energy), while the intent term remains active to allow proactive reservation for imminent actions (e.g., "I will share my slides" even before the UI action happens). The attention term captures receiver-side UI emphasis.

Handling role–mode uncertainty. Given the MLLM posterior $\boldsymbol{\pi}_i(t)$, IVA can compute the final score in two equivalent ways:

$$w_{ij}(t) = \sum_{g=1}^{G} \pi_i^{(g)}(t)\, w_{ij}^{(g)}(t) \quad \text{(mixture)}, \quad (2)$$

or, for a simpler hard selection, $w_{ij}(t) = w_{ij}^{(\hat{g})}(t)$ where $\hat{g} = \arg\max_g \pi_i^{(g)}(t)$.

Normalization across competing senders. For each receiver j, priorities are normalized across all incoming streams to form semantic weights:

$$\tilde{w}_{ij}(t) = \frac{\left[w_{ij}(t)\right]_+}{\varepsilon + \sum_{k \neq j} \left[w_{kj}(t)\right]_+}, \quad (3)$$

where $[x]_+ \triangleq \max\{x, 0\}$ and $\varepsilon > 0$ prevents division-by-zero. If the denominator is effectively zero (all scores non-positive), IVA falls back to uniform weights for receiver j.

4.3 QoE Formulation

We define QoE at the stream, receiver, and system levels.

(1) Stream-level quality. Let $q_{ij}(t)$ denote the perceived quality contribution of stream $i \to j$ during interval t. We adopt a tractable signal-level model:

$$q_{ij}(t) = \alpha\, b_{ij}(t) - \beta\, d_{ij}(t), \quad (4)$$

where $\alpha, \beta > 0$ scale the relative importance of bitrate and latency.

Latency model. Define the sender uplink load and receiver downlink load as

$$b_i^{\text{up}}(t) \triangleq \sum_{k \neq i} b_{ik}(t), \qquad b_j^{\text{down}}(t) \triangleq \sum_{k \neq j} b_{kj}(t). \quad (5)$$

We approximate end-to-end latency as the sum of uplink transmission time, downlink transmission time, and IVA's processing/actuation overhead:

$$d_{ij}(t) = \Delta t \left(\frac{b_i^{\text{up}}(t)}{B_i^{\text{up}}(t)} + \frac{b_j^{\text{down}}(t)}{B_j^{\text{down}}(t)}\right) + L_{\text{sem}}(t), \quad (6)$$

where $L_{\text{sem}}(t)$ captures the additional delay attributable to semantic processing and control actuation in IVA (e.g., intent inference and scheduling overhead). This term is empirically small in our prototype (tens of milliseconds), but we include it explicitly to quantify IVA's latency cost.

To penalize temporal instability, we define the quality variation

$$\Delta q_{ij}(t) \triangleq \left|q_{ij}(t) - q_{ij}(t-1)\right|. \quad (7)$$

(2) Receiver-level QoE. Receiver j's QoE at time t is a semantic-weighted sum of stream qualities with an instability penalty:

$$QoE_j(t) = \sum_{\substack{i=1 \\ i \neq j}}^{n} \tilde{w}_{ij}(t)\, q_{ij}(t) - \lambda \sum_{\substack{i=1 \\ i \neq j}}^{n} \Delta q_{ij}(t), \quad (8)$$

where $\lambda > 0$ penalizes rapid changes in perceived quality.

(3) System-level QoE. System QoE aggregates over receivers:

$$QoE(t) = \sum_{j=1}^{n} \mu_j\, QoE_j(t), \quad (9)$$

where $\mu_j > 0$ weights receiver importance (uniform in our evaluation unless stated otherwise).

Constraints. IVA operates under two system constraints.
(1) Bandwidth constraints.

$$\sum_{j=1}^{n} b_{ij}(t) \le B_i^{\text{up}}(t), \qquad \sum_{i=1}^{n} b_{ij}(t) \le B_j^{\text{down}}(t), \tag{10}$$

for all t, with the convention $b_{ii}(t) = 0$.
(2) Latency constraint. To preserve interactivity, the (time-averaged) end-to-end delay must remain below a target t_l:

$$\frac{1}{T}\sum_{t=1}^{T} d_{ij}(t) \le t_l, \qquad \forall\, i \ne j. \tag{11}$$

Optimization objective. Activity and control semantics determine *who matters now*; intent semantics anticipate *who will matter next*. IVA allocates bitrate to maximize time-averaged system QoE:

$$\max_{\{b_{ij}(t)\}} \quad \frac{1}{T}\sum_{t=1}^{T} QoE(t) \quad \text{s.t.} \quad (10),\ (11). \tag{12}$$

In practice, IVA solves this objective online in a receding-horizon manner: at each interval, it uses current semantic weights $\tilde{w}_{ij}(t)$ and network measurements to compute $b_{ij}(t)$ while keeping feasibility with respect to the instantaneous bandwidth limits, and uses the instability penalty to stabilize allocations over time.

5 RL-based Solution

The key challenge in IVA is that the mapping from semantics to priority, i.e., $\Theta^{(g)} = \{\boldsymbol{\theta}_{\text{act}}^{(g)}, \boldsymbol{\theta}_{\text{int}}^{(g)}, \kappa^{(g)}\}$ in Eq. (1), is inherently *context dependent*. Presentation mode favors shared-content protection, while free discussion prioritizes low latency and rapid reallocation. Because hard-coded trade-offs are brittle, IVA learns role–mode–specific priorities via reinforcement learning (RL) on trace-driven simulations.

Two-level control. IVA decouples learning from feasibility: RL *learns* semantic priorities $w_{ij}(t)$ and normalized weights $\tilde{w}_{ij}(t)$ via Eq. (1) and Eq. (3), while a constrained bitrate allocator computes feasible $\{b_{ij}(t)\}$ subject to Eq. (10) and Eq. (11). This separation learns *what to protect* while ensuring *how to allocate* always respects network constraints.

MDP formulation. IVA learning is formulated as an episodic Markov decision process (MDP) over trace-driven conferencing segments of horizon T.

State. At time t, the state aggregates semantic features, UI controls, network context, and prior-step quality:

$$s_t = \Big(\{\mathbf{x}_{ij}^{\text{act}}(t), \mathbf{x}_{ij}^{\text{int}}(t), m_i(t), a_{ij}(t)\}_{i\ne j}, \\ \{B_i^{\text{up}}(t)\}_i,\ \{B_j^{\text{down}}(t)\}_j, \{q_{ij}(t-1)\}_{i\ne j}\Big). \tag{13}$$

Action. For each role–mode g, the policy outputs unnormalized priority scores $w_{ij}^{(g)}(t)$ via Eq. (1), which are combined using Eq. (2) or hard selection, normalized to $\tilde{w}_{ij}(t)$ via Eq. (3), and passed to the constrained bitrate allocator.

Environment. Given $\tilde{w}_{ij}(t)$, a trace-driven emulator computes feasible bitrates $b_{ij}(t)$ under Eq. (10), simulates delivery, and returns the next state s_{t+1}.

Reward. The reward is the instantaneous system QoE (Eq. (9)), regularized to penalize aggressive prioritization:

$$r_t = \frac{1}{Q_{\text{ref}}} QoE(t) - \lambda_\Theta \sum_{g=1}^{G} \left\| \Theta^{(g)} \right\|_2^2, \tag{14}$$

where Q_{ref} normalizes rewards and $\lambda_\Theta > 0$ controls regularization.

Policy optimization and adaptation. IVA adopts the structured gated–bias parameterization in Eq. (1) for interpretability and stability. Exploration adds Gaussian noise to priority scores, $\hat{w}_{ij}^{(g)}(t) = w_{ij}^{(g)}(t) + \epsilon_{ij}^{(g)}(t)$, $\epsilon_{ij}^{(g)}(t) \sim \mathcal{N}(0, \sigma_g^2)$, and computing $\tilde{w}_{ij}(t)$ via Eq. (2) and Eq. (3), yielding a continuous stochastic policy $\pi_\Theta(\hat{w}(t) \mid s_t)$.

Learning algorithm. IVA trains its policy using proximal policy optimization (PPO), with a learned critic $V_\phi(s)$ estimating advantages A_t from discounted returns. Policy parameters $\Theta = {\Theta^{(g)}}_{g=1}^{G}$ maximize the standard clipped PPO objective, while the critic minimizes squared TD error. In implementation, we use clipped likelihood-ratio updates to prevent abrupt policy shifts under rapidly changing traces, and estimate advantages with a generalized-advantage-style return baseline for lower-variance gradients.

Offline training and safe online adaptation. IVA is trained offline on trace-driven simulations, guided by MLLM-inferred role–mode posteriors. At runtime, the corresponding $\Theta^{(g)}$ is selected to compute semantic priorities via Eq. (1), which drive the constrained bitrate allocator to produce $b_{ij}(t)$. Optional small trust-region PPO updates enable online adaptation to session dynamics, while bandwidth and latency constraints are always enforced to ensure safety.

6 Experiment Evaluation

Experiment setup. We implement a WebRTC-based prototype supporting multi-party audio, video, and UI events. IVA is evaluated using FCC broadband traces [19] and HSDPA mobile traces [20] by replaying conferencing workloads and measuring QoE and end-to-end latency ($d_{ij}(t)$). We use a discrete-time simulation with $\Delta t = 100$ ms, $T = 60$ s, and $H = 2$ s. Following [21], we set $\alpha = 2$, $\beta = 1$, initialize $\mu \sim \mathcal{N}^+(1.0,\ 0.3^2)$, model semantic delay as $L_{\text{sem}}(t)$, fix $t_l = 400$ ms [22], and set $\gamma = 0.95$, $\lambda = 0.5$. A PPO agent is trained for 2000 episodes with learning rate decaying from 5×10^{-5} to 1×10^{-5}, clip ratio 0.2, value loss 0.8, entropy

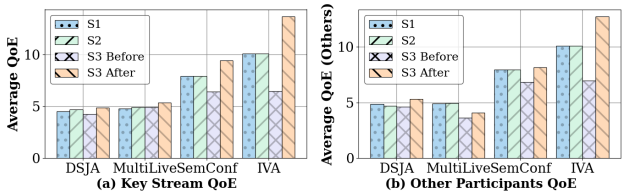

Figure 3: QoE comparison across scenarios.

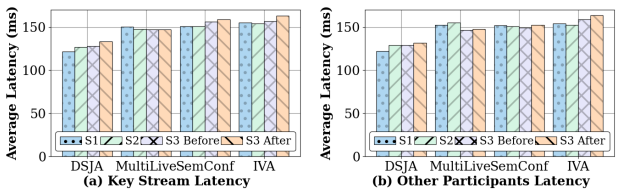

Figure 4: Latency comparison across scenarios.

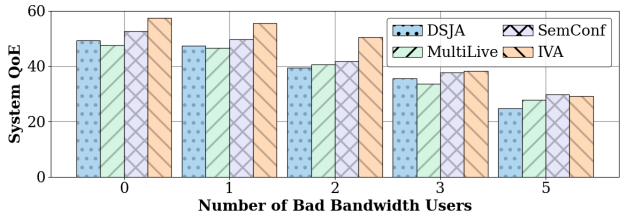

Figure 5: QoE under degraded network conditions.

0.08, and 4 update epochs per step. We evaluate three five-participant scenarios:

S1: Presentation with Screen Sharing. A primary speaker announces screen sharing at 10–20 s and then begins sharing, while other participants remain mostly muted. The goal is to preserve high QoE for audio and shared content.

S2: Multi-Party Discussion. Participants engage in short, balanced speaking turns without a dominant presenter or screen sharing, stressing the need for rapid bitrate reallocation under frequent floor changes.

S3: Mode Transition. The session transitions from presentation to interactive Q&A, shifting priorities from presenter-centric delivery to low-latency interaction.

Performance evaluation. We compare IVA with MultiLive [11], DSJA [12], and SemConf [21]. We report QoE for the key stream (primary speaker in S1/S3; all participants in S2) and average QoE of other participants. Fig. 3 shows IVA outperforms all baselines in S1 and S2, and in S3 it recovers key-stream and audience QoE after the presentation-to-Q&A transition by quickly re-centering protection from presenter-dominant delivery to interactive exchange. Fig. 4 shows IVA adds about 25 ms latency over DSJA while remaining below 165 ms across scenarios, within real-time thresholds [22].

Fig. 5 shows that IVA outperforms signal-driven ABR schemes when one or more users are degraded. It also outperforms SemConf when one to three users are affected and

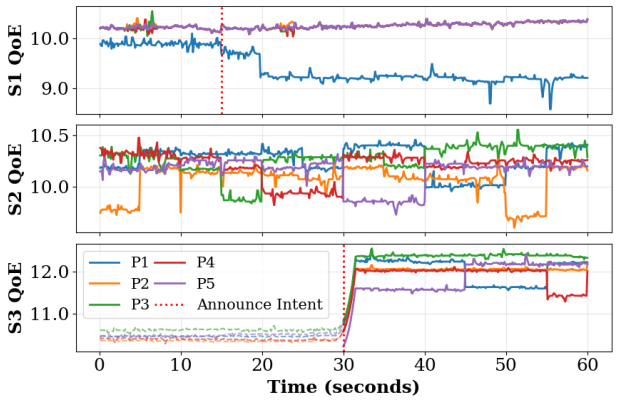

Figure 6: User-level QoE comparison over time.

achieves comparable QoE when all five users are degraded, demonstrating robustness under adverse conditions.

Fig. 6 shows per-user QoE evolution across scenarios. In S1, around 15 s, IVA detects announced screen sharing and proactively reallocates bitrate, slightly reducing P1's QoE while improving audience QoE. In S2, with short balanced turns, IVA dynamically adjusts priorities as speakers change, lowering the active speaker's incoming QoE while improving others'. In S3, after the transition from presentation to interactive Q&A, IVA shifts priorities toward low-latency interaction and increases QoE for all participants.

To isolate intent semantics, we compare IVA with **IVA w/o Intent**, which uses only activity and control semantics. Intent semantics improve primary-speaker QoE by 14.0% and 36.5% in S1/S3, and improve other participants by 7.97% and 29.5%. In S2, IVA improves QoE by 14.0% for all users. Across scenarios, intent inference adds under 30 ms latency, keeping total delay below 165 ms and within practical thresholds [22].

7 Concluding Remarks

Conversational text in video meetings often signals intent before quality-critical events, yet current conferencing controllers remain reactive to network measurements. IVA treats transcripts as control input: streaming ASR and an MLLM infer activity, control, and near-future intent semantics, a gated–bias model maps them to per-stream priorities, and a constrained allocator orchestrates bitrate under uplink, downlink, and latency budgets. This closes the loop between language understanding and media adaptation beyond signal-only ABR and prior semantic-aware designs, and experiments show improved QoE around impending transitions.

Acknowledgments

This work was supported by the UGC General Research Fund no. 17209822 and the Innovation and Technology Commission Fund no. ITS/383/23FP from Hong Kong.

References

[1] Yili Jin, Junhua Liu, Kaiyuan Hu, and Fangxin Wang. A networking perspective of volumetric video service: Architecture, opportunities and case study. *IEEE Network*, 2024.

[2] I. Sandvine. The Global Internet Phenomena Report. https://www.applogicnetworks.com/phenomena, March 2024. Accessed: January 16, 2026.

[3] Jian Wu, Yashesh Gaur, Zhuo Chen, Long Zhou, Yimeng Zhu, Tianrui Wang, Jinyu Li, Shujie Liu, Bo Ren, Linquan Liu, et al. On decoder-only architecture for speech-to-text and large language model integration. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE, 2023.

[4] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

[5] Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023.

[6] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

[7] Fireflies.ai | The #1 AI Notetaker for Your Meetings. https://fireflies.ai, 2025. Accessed: January 16, 2026.

[8] AI note taker: Real-time meeting assistant. https://www.zoom.com/en/products/ai-assistant/features/ai-note-taking/, 2025. Accessed: January 16, 2026.

[9] Transcription Made Easy with Otter.ai. https://get.otter.ai/otter_ai_chatgpt/, 2025. Accessed: January 16, 2026.

[10] Branislav Sredojev, Dragan Samardzija, and Dragan Posarac. Webrtc technology overview and signaling solution design and implementation. In *2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)*, pages 1006–1009. IEEE, 2015.

[11] Ziyi Wang, Yong Cui, Xiaoyu Hu, Xin Wang, Wei Tsang Ooi, Zhen Cao, and Yi Li. Multilive: Adaptive bitrate control for low-delay multi-party interactive live streaming. *IEEE/ACM Transactions on Networking*, 30(2):923–938, 2021.

[12] Dayou Zhang, Hao Zhu, Kai Shen, Dan Wang, and Fangxin Wang. Dsja: Distributed server-driven joint route scheduling and streaming adaptation for multi-party realtime video streaming. *IEEE Transactions on Mobile Computing*, 23(7):7680–7694, 2023.

[13] Zhenzi Weng, Zhijin Qin, Xiaoming Tao, Chengkang Pan, Guangyi Liu, and Geoffrey Ye Li. Deep learning enabled semantic communications with speech recognition and synthesis. *IEEE Transactions on Wireless Communications*, 22(9):6227–6240, 2023.

[14] Huiqiang Xie, Zhijin Qin, Geoffrey Ye Li, and Biing-Hwang Juang. Deep learning enabled semantic communication systems. *IEEE transactions on signal processing*, 69:2663–2675, 2021.

[15] Zhenzi Weng and Zhijin Qin. Semantic communication systems for speech transmission. *IEEE Journal on Selected Areas in Communications*, 39(8):2434–2444, 2021.

[16] Ziliang Zhou, Shilian Zheng, Jie Chen, Zhijin Zhao, and Xiaoniu Yang. Speech semantic communication based on swin transformer. *IEEE Transactions on Cognitive Communications and Networking*, 10(3):756–768, 2023.

[17] AI Model Evaluation Platform. https://www.shengwang.cn/duihua/benchmark/, 2025. Accessed: January 16, 2026.

[18] Jia He, Mostafa Ammar, Ellen Zegura, and Emir Halepovic. Qoe metrics for interactivity in video conferencing applications: definition and evaluation methodology. In *Proceedings of the 15th ACM Multimedia Systems Conference*, pages 178–189, 2024.

[19] Measuring Broadband Raw Data Releases - Fixed. https://www.fcc.gov/oet/mba/raw-data-releases, 2025. Accessed: January 6, 2026.

[20] Haakon Riiser, Paul Vigmostad, Carsten Griwodz, and Pål Halvorsen. Commute path bandwidth traces from 3g networks: Analysis and applications. In *Proceedings of the 4th ACM Multimedia Systems Conference*, pages 114–118, 2013.

[21] Xize Duan, Yili Jin, Lei Zhang, and Fangxin Wang. Semconf: A system for multiparty semantic video conferencing. In *Proceedings of the 35th Workshop on Network and Operating System Support for Digital Audio and Video*, pages 71–77, 2025.

[22] Mohammad H Hajiesmaili, Lok To Mak, Zhi Wang, Chuan Wu, Minghua Chen, and Ahmad Khonsari. Cost-effective low-delay design for multiparty cloud video conferencing. *IEEE Transactions on Multimedia*, 19(12):2760–2774, 2017.