BEYOND REACTIVE VISUAL ASSISTANCE: A REAL-TIME SCENE-INTERACTION AGENT FOR BLIND AND LOW-VISION PEOPLE

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual assistance for blind and low-vision (BLV) people remains largely query-driven: users must ask about cues they may not know exist. Real-time BLV assistance from continuous first-person video instead requires an agent that decides when evidence warrants speech, avoids redundant notifications, and delivers scene-appropriate guidance while that evidence remains actionable. We introduce a BLV-informed proactive scene-interaction dataset by densely annotating 439 multi-scene first-person video segments from BLV users' daily lives, yielding 7,482 timestamp-aligned PUSH/NO-PUSH decisions along continuous trajectories and assistance messages for 1,794 PUSH decisions. We present BLV-SceneAgent, a real-time observe–decide–act–commit agent built on a streaming VLM. Its Temporal Contextual Intervention (TCI) policy integrates the native response margin with a causal recurrent belief, signed state change, and a committed-notification ledger, and is refined under the agent's predicted action history. NO-PUSH terminates at the action boundary, whereas PUSH reuses the multimodal cache for content generation. On a segment-disjoint test set, TCI improves steady-state macro-F1 by 6.3 points over calibrated SFT, with a P95 action-boundary latency of only 478.8 ms. A live study across four settings shows that the agent surpasses baselines in usefulness, trust, and reduced annoyance. We will publicly release our dataset and model to support further research on proactive visual assistance for blind and low-vision people.

1 INTRODUCTION

Reactive visual assistance is least effective when a user cannot formulate a query about an unseen cue. Blind and low-vision (BLV) people seek information about text, objects, appearance, activities, spatial layout, and object state—well beyond obstacle detection (Brady et al., 2013; Gurari et al., 2018). Yet most systems wait for a captured image or clip and an explicit question (Bigham et al., 2010). This reactive contract cannot reliably surface a vehicle entering the walking path, debris revealed during cleaning, residue left on an object after washing, or a directed social gesture before the opportunity to act has passed.

Persistent assistance cannot simply narrate every frame. Speech competes with environmental audio, consumes attention, and becomes disruptive when it repeats already communicated information. Real-time scene interaction therefore couples three capabilities: *actionable latency*, so a decision arrives while evidence is useful; *temporal selectivity*, so meaningful change is separated from repetition; and *BLV-oriented generation*, so speech is concise, situated, and actionable. Because each notification changes what remains worth saying, this is a sequential agent problem rather than framewise classification followed by captioning. Figure 1 contrasts this closed-loop interaction with reactive visual description.

Existing resources capture authentic visual questions, teachable recognition, timestamp-restricted first-person video question answering (VideoQA), and live description (Gurari et al., 2018; Massiceti et al., 2021; Xiao et al., 2025; Chang et al., 2024). Streaming VLMs further enable temporally aligned responses (Chen et al., 2024; Qian et al., 2025; Ding et al., 2025; Wang et al., 2025). What remains missing is chronological supervision for intervention: whether current evidence warrants

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

speech given what the system has already communicated, and what assistance should be delivered when it does.

We address this gap with a multi-scene proactive annotation layer over 439 first-person EgoBlind segments (Xiao et al., 2025). Five sighted–BLV annotator dyads produced 7,482 decisions along continuous trajectories and 1,794 timestamp-aligned PUSH messages. The paired protocol combines detailed visual transcription with BLV lived experience: BLV annotators determine both whether assistance is warranted and what would be useful to hear. Consequently, the same visible object can be labeled PUSH when newly relevant and NO-PUSH after its information has been communicated, directly supervising the distinction between visual salience and intervention utility.

We present BLV-SceneAgent, a persistent observe–decide–act–commit agent learned in two stages. First, action-first SFT adapts a streaming VLM to chronological BLV supervision, jointly learning the SILENCE/RESPONSE decision boundary and concise, scene-appropriate assistance. At every tick, the adapted backbone produces a causal action state and response margin while retaining a reusable multimodal key–value (KV) cache for conditional continuation (Yao et al., 2026). Second, Temporal Contextual Intervention (TCI) turns this per-tick representation into a closed-loop policy: a recurrent belief summarizes causal history, signed state change modulates intervention utility, and a committed-notification ledger suppresses redundant speech. Self-history refinement exposes the policy to its own executed actions. SILENCE terminates at the action boundary, whereas RESPONSE continues from the same cache; persistent state is committed only after the selected action succeeds. This transactional path jointly preserves real-time execution, temporal selectivity, and BLV-oriented content generation.

Our contributions are: (1) a BLV-informed, multi-scene proactive annotation layer with timestamp-aligned supervision for both whether to intervene and what assistance to provide; (2) a real-time scene-interaction agent that couples a BLV-adapted streaming VLM, a committed-notification ledger, and conditional generation in an observe–decide–act–commit loop; and (3) TCI, a causal intervention policy that learns dynamic utility, signed state-change modulation, and sign-constrained repetition suppression from chronological point labels, then refines the policy under its own predicted action history.

2 RELATED WORK

2.1 VISUAL ASSISTANCE FOR BLV PEOPLE

Visual assistance for blind and low-vision (BLV) people has progressed from on-demand access toward personalized and continuous support. Studies of everyday visual questions reveal needs that extend beyond obstacle avoidance to reading, object identity, appearance, activity, spatial layout, and object state (Stangl et al., 2020; 2021). Human-in-the-loop services made such information accessible through conversational assistance (Lasecki et al., 2013), while subsequent vision systems broadened automated support. ORBIT, for example, recognizes personally relevant objects from BLV-captured videos (Massiceti et al., 2021); analyses of vision–language representations further show that models trained on web imagery do not transfer uniformly to BLV visual data (Massiceti et al., 2024).

More recent work moves from isolated queries toward assistance embedded in an ongoing activity. WorldScribe studies context-aware live description (Chang et al., 2024), and ICanSee examines real-time VideoLLMs for visual-impairment assistance (Zhang et al., 2025b). This progression reduces the delay between perception and support, but it largely retains a request–response or continuous-description contract. Proactive assistance requires a different capability: the system must decide whether a newly observed change is worth interrupting the user for, conditioned on both the current activity and what has already been communicated. We formulate this missing capability as chronological intervention learning, jointly supervising informative speech and deliberate silence from BLV-informed judgments.

2.2 STREAMING MULTIMODAL LARGE MODELS

Streaming multimodal large models replace bounded-clip inference with causal, incremental processing of an unfolding visual stream. VideoLLM-online established temporally aligned online

video dialogue (Chen et al., 2024). Subsequent systems made the interaction loop more structured: Dispider decomposes perception, decision, and reaction (Qian et al., 2025); StreamMind gates language reasoning on detected events (Ding et al., 2025); MMDuet interleaves video and text for time-sensitive interaction (Wang et al., 2025); and real-time interaction models combine causal action prediction with conditional response generation (Yao et al., 2026). Collectively, these advances provide the architectural basis for low-latency perception and response without repeatedly encoding the entire video history.

A complementary line of work extends streaming response toward proactive behavior. Egocentric assistants have been trained with synthesized interactions and multi-horizon memory (Zhang et al., 2025a; Gong et al., 2026); Proact-VL explicitly targets low-latency proactive companions (Yan et al., 2026); and EgoSpeak models speech initiation in online egocentric conversation (Kim et al., 2025). These methods establish that a streaming model can initiate rather than merely answer. BLV scene interaction, however, also requires *notification-conditioned selectivity*: after an observation has been verbalized, its utility should diminish unless its state or risk materially changes. BLV-SceneAgent closes this loop by coupling streaming multimodal inference with an intervention policy whose recurrent state is updated by executed notifications, thereby learning both timely assistance and contextually appropriate silence.

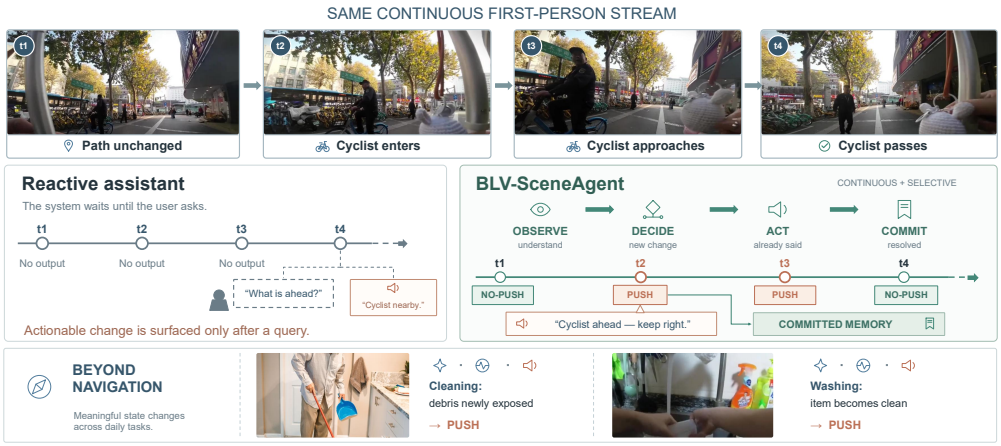


Figure 1: From reactive description to proactive scene interaction. Reactive assistance waits for a query, whereas BLV-SceneAgent continuously selects whether to speak. Given the same stream, the agent intervenes when a cyclist enters the user's path and commits the notification. The same intervention policy applies to actionable state changes during cleaning and washing.

3 METHOD

3.1 BLV-SCENEAGENT: CAUSAL OBSERVE–DECIDE–ACT–COMMIT

We model proactive assistance as sequential intervention over a continuous first-person stream. Let $\tau_1 < \cdots < \tau_T$ denote ordered decision times. At tick t, the agent receives only the causal clip $o_t = X_{\tau_t - L:\tau_t}$ and a bounded serialization B_t of actions executed strictly before τ_t. It selects an intervention $d_t \in \{0, 1\}$ and, only when $d_t = 1$, an assistance message y_t:

$$a_t = \begin{cases} \textsc{Silence}, & d_t = 0, \\ \textsc{Speak}(y_t), & d_t = 1. \end{cases} \tag{1}$$

We map $d_t = 0/1$ to the streaming backbone's native $\langle\textsc{Silence}\rangle/\langle\textsc{Response}\rangle$ tokens and refer to the executed actions as NO-PUSH/PUSH. Supervision consists of chronological trajectories $\mathcal{D} = \{(o_t, B_t^\star, d_t^\star, y_t^\star)\}_{t=1}^{T}$, where $y_t^\star$ exists only at positive ticks and the current target is absent from both inputs. At deployment, future state is determined by the agent's executed predictions rather than future annotations. The task therefore couples timely intervention, deliberate silence, and conditional assistance generation in one causal policy.

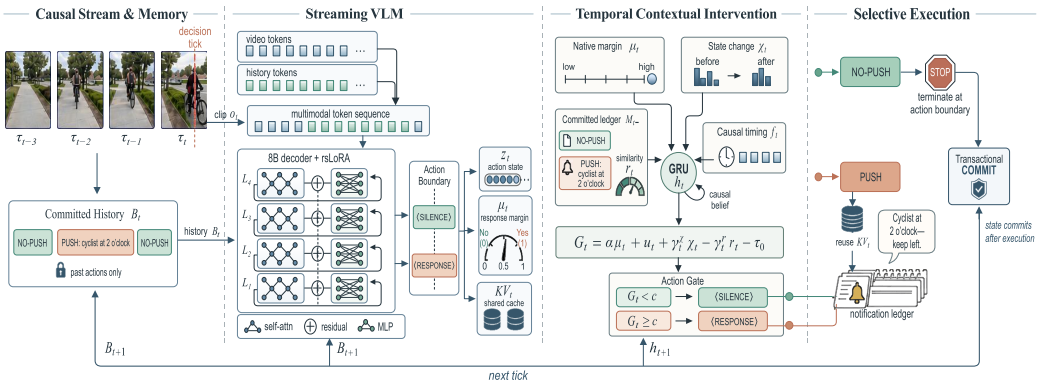


Figure 2: BLV-SceneAgent as a causal observe–decide–act–commit agent. One prefill exposes the action boundary; TCI decides from recurrent belief and the committed-notification ledger. NO-PUSH stops, whereas PUSH continues the action-first BLV SFT decoder from the same KV cache; state commits only after execution.

Figure 2 illustrates the complete agent. Before tick t, its private state is

$$s_{t^-} = (h_{t-1}, \mathcal{M}_{t^-}, v_{t-1}, \vartheta_t), \tag{2}$$

where h_{t-1} is a causal recurrent belief, $\mathcal{M}_{t^-}$ is a bounded committed-notification ledger of executed PUSH actions, v_{t-1} is the previous normalized action state, and ϑ_t stores past timestamps and the previous executed action. A streaming VLM performs one multimodal prefill:

$$(z_t, \mu_t, \mathrm{KV}_t) = \Phi(o_t, B_t), \qquad \mu_t = \ell_t^{\textsc{Response}} - \ell_t^{\textsc{Silence}}, \tag{3}$$

where $z_t \in \mathbb{R}^D$ is the penultimate text-decoder state at the action boundary, μ_t is the native response margin, and KV_t is the reusable visual-language cache. TCI reads (z_t, μ_t, s_{t^-}), proposes an action, and changes no persistent controller state until model-side token generation succeeds. SILENCE requires no autoregressive continuation; RESPONSE continues from KV_t. After the selected token sequence is generated successfully, the controller state—and the ledger for RESPONSE—is committed, and the output tokens are detokenized and passed to the downstream speech interface. This transactional update prevents an abandoned response from suppressing later assistance.

3.2 Action-First BLV Adaptation

The VLM is trained with two native output forms:

$$\langle \textsc{Silence} \rangle \quad \text{or} \quad \langle \textsc{Response} \rangle \| y_t^\star. \tag{4}$$

We retain the 8B checkpoint in BF16 and train rank-64 rank-stabilized low-rank adapters (Hu et al., 2022). The vision tower and multimodal projector remain unchanged; adapters target language-model linear layers. Let $r_t^\star = \langle \textsc{Silence} \rangle$ for a negative tick and $r_t^\star = \langle \textsc{Response} \rangle \| y_t^\star$ for a positive tick. Standard teacher-forced causal-language-model fine-tuning minimizes

$$\mathcal{L}_{\mathrm{SFT}} = -\sum_t \sum_{m=1}^{|r_t^\star|} \log p_\Phi(r_{t,m}^\star \mid o_t, B_t^\star, r_{t,<m}^\star). \tag{5}$$

SFT operates on timestamp-indexed causal examples. At every tick, the first action token receives supervision; message tokens are optimized only for positive decisions. The English input combines a materialized 3-second, 2-fps causal clip (maximum side 384 pixels) with $B_t^\star$, which contains the four most recent actions and up to five earlier PUSH records, all strictly before the current tick.

3.3 Temporal Contextual Intervention

We project the action state and update a compact gated recurrent unit (GRU) belief (Cho et al., 2014):

$$e_t = \mathrm{SiLU}(W_p\, \mathrm{LN}(z_t)), \tag{6}$$

$$h_t = \mathrm{GRU}([e_t; \tanh(\mu_t / s_\mu); f_t], h_{t-1}), \tag{7}$$

where f_t contains seven bounded, causal features: cold start, whether the previous committed action was PUSH, elapsed time since the previous tick, whether any PUSH has been committed, elapsed time since that action, latent-state change, and visual repetition. The recurrent state predicts a residual intervention utility u_t, a latent intent q_t, and two contextual gain logits:

$$u_t = w_u^\top \mathrm{SiLU}(W_u \,\mathrm{LN}(h_t)), \tag{8}$$

$$q_t = \mathrm{norm}(W_q \,\mathrm{LN}(h_t)), \tag{9}$$

$$[g_t^\chi, g_t^r] = W_g \,\mathrm{LN}(h_t). \tag{10}$$

Here norm denotes ℓ_2 normalization.

Let $v_t = z_t/\|z_t\|_2$. State change is robustly scaled using training-set quantiles of $1 - \langle v_t, v_{t-1}\rangle$:

$$\chi_t = \mathrm{clip}\left(\frac{1 - \langle v_t, v_{t-1}\rangle - Q_{0.1}^\chi}{Q_{0.9}^\chi - Q_{0.1}^\chi}, 0, 1\right), \tag{11}$$

where $Q_{0.1}^\chi$ and $Q_{0.9}^\chi$ are the training-set 10th and 90th percentiles, and $\chi_1 = 0$. The committed-notification ledger contains at most K tuples (τ_j, q_j, v_j). Semantic repetition and visual repetition are

$$r_t^{\mathrm{sem}} = \max_{j \in \mathcal{M}_{t^-}} \exp\left(-\frac{\tau_t - \tau_j}{\kappa}\right) \frac{1 + \langle q_t, q_j\rangle}{2}, \tag{12}$$

$$r_t^{\mathrm{vis}} = \mathrm{RobustScale}\left(\max_{j \in \mathcal{M}_{t^-}} \langle v_t, v_j\rangle\right). \tag{13}$$

Both repetition terms are zero when $\mathcal{M}_{t^-}$ is empty; until the first committed PUSH, elapsed-since-PUSH is set to its capped value. When the current state changes sharply from the preceding tick, similarity to an older committed notification is less reliable evidence that the present information is redundant. We therefore define

$$r_t = \max(r_t^{\mathrm{sem}}, r_t^{\mathrm{vis}})(1 - \chi_t). \tag{14}$$

The repetition gain is non-negative, whereas the change gain is bounded and signed:

$$\gamma_t^\chi = \gamma_{\max} \tanh(g_t^\chi), \quad \gamma_t^\chi \in [-\gamma_{\max}, \gamma_{\max}], \tag{15}$$

$$\gamma_t^r = \mathrm{softplus}(g_t^r). \tag{16}$$

The adjusted action score is

$$G_t = \alpha\mu_t + u_t + \gamma_t^\chi \chi_t - \gamma_t^r r_t - \tau_0, \qquad \alpha > 0. \tag{17}$$

Here τ_0 is a learned scalar base threshold and α is a learned positive scale for the native VLM margin. We use $s_\mu = 4$, $\gamma_{\max} = 0.5$, $\kappa = 20$ seconds, and $K = 8$; each $\mathrm{RobustScale}$ uses training-only 10th/90th percentiles followed by clipping to $[0, 1]$. The agent executes PUSH if and only if t is the mandatory cold start or $G_t \geq c$, where the single global offset c is selected on validation segments. The signed gain lets state change either increase or decrease intervention utility as a function of the recurrent context, accommodating both meaningful transitions and uninformative egocentric motion. In contrast, the sign-constrained repetition gain is monotone: greater similarity to a committed notification can only suppress the current score. A single global threshold is shared across scenes; scene-dependent intervention frequency emerges from the native margin and causal recurrent belief learned from the annotated trajectories, without a hand-designed rate quota.

3.4 LEARNING THE INTERVENTION POLICY

After SFT, Φ is frozen and TCI is trained on chronological action-state trajectories. Let $\mathcal{I}$ exclude the mandatory first tick of each annotated segment, and let N^+ and N^- be its positive and negative counts. Class imbalance is handled with $w^+ = \min(4, N^-/N^+)$. The principal loss is weighted binary cross-entropy,

$$\mathcal{L}_{\mathrm{point}} = \frac{1}{|\mathcal{I}|} \sum_{t \in \mathcal{I}} \mathrm{BCE}(G_t, d_t^\star; w^+). \tag{18}$$

Let $p_t = \sigma(G_t)$, $\mathcal{T} = \{t \in \mathcal{I} : d_t^\star \neq d_{t-1}^\star\}$, and $\mathcal{S} = \{t \in \mathcal{I} : d_t^\star = d_{t-1}^\star\}$. We use

$$\mathcal{L}_{\mathrm{tr}} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \mathrm{BCE}(G_t, d_t^\star; w^+), \tag{19}$$

$$P_s = \frac{\sum_{t \in \mathcal{I}} p_t d_t^\star}{\max(\sum_{t \in \mathcal{I}} p_t, \epsilon)}, \quad R_s = \frac{\sum_{t \in \mathcal{I}} p_t d_t^\star}{\max(\sum_{t \in \mathcal{I}} d_t^\star, \epsilon)}, \tag{20}$$

$$\mathcal{L}_{\mathrm{softF1}} = 1 - \frac{2 P_s R_s}{\max(P_s + R_s, \epsilon)}, \tag{21}$$

$$\mathcal{L}_{\mathrm{stab}} = \frac{1}{|\mathcal{S}|} \sum_{t \in \mathcal{S}} (p_t - p_{t-1})^2, \tag{22}$$

with $\epsilon = 10^{-6}$; an empty auxiliary set contributes zero. We also rank positive above negative ticks from the same annotated segment. Let $\mathcal{I}_v \subseteq \mathcal{I}$ contain the eligible ticks of segment v; the outer expectation averages segments containing both labels:

$$\mathcal{L}_{\mathrm{rank}} = \mathbb{E}_v \, \mathbb{E}_{\substack{i,j \in \mathcal{I}_v \\ d_i^\star = 1,\ d_j^\star = 0}} \mathrm{softplus}(-(G_i - G_j)), \tag{23}$$

Intent prediction is causal at inference: q_t is computed from the current action state and recurrent controller state. During training, non-cold-start PUSH messages provide semantic targets through a frozen English all-MiniLM-L6-v2 sentence encoder (Reimers & Gurevych, 2019) and a PCA projection fitted on training messages, yielding normalized targets $q_t^\star$. We minimize

$$\mathcal{L}_{\mathrm{int}} = \frac{1}{N^+} \sum_{\substack{t \in \mathcal{I} \\ d_t^\star = 1}} (1 - \langle q_t, q_t^\star \rangle). \tag{24}$$

We also use the negative-tick repetition penalty

$$\mathcal{L}_{\mathrm{rep}} = \mathbb{E}_{\substack{t \in \mathcal{I} \\ d_t^\star = 0}} [\sigma(G_t) \, \mathrm{sg}(r_t^{\mathrm{sem}})]. \tag{25}$$

Here σ is the logistic sigmoid and sg stops gradients. An auxiliary category head $\pi_t = \mathrm{softmax}(W_s \, \mathrm{LN}(h_t))$ is supervised by cross-entropy $\mathcal{L}_{\mathrm{scene}}$ on all valid training ticks, regularizing the recurrent representation. Runtime decisions remain governed by the category-independent score in Equation 17; every other controller loss uses $\mathcal{I}$. The selected training objective is

$$\begin{aligned} \mathcal{L}_{\mathrm{TCI}} = \mathcal{L}_{\mathrm{point}} &+ .05\mathcal{L}_{\mathrm{tr}} + .20\mathcal{L}_{\mathrm{rank}} \\ &+ .10\mathcal{L}_{\mathrm{softF1}} + .005\mathcal{L}_{\mathrm{stab}} + .05\mathcal{L}_{\mathrm{int}} \\ &+ .02\mathcal{L}_{\mathrm{rep}} + .05\mathcal{L}_{\mathrm{scene}}. \end{aligned} \tag{26}$$

TCI is optimized in two supervised stages. The first stage performs behavior cloning under gold preceding actions. Self-history refinement then rolls the current policy chronologically through each training segment with $c = 0$ and uses its executed predictions to construct the next epoch's action history and committed-notification ledger. Cached VLM states remain conditioned on the fixed prior serialization, isolating controller adaptation. This dataset-aggregation-style replay exposes the policy to its own false positives and missed notifications, reducing training–deployment history mismatch without additional annotation (Ross et al., 2011); the final c is calibrated on validation trajectories.

3.5 SAME-PREFILL ASSISTANCE GENERATION

The first native action token is both the decision boundary and the prefix for content generation. When TCI selects PUSH, the agent resumes the action-first SFT model from the already-computed multimodal cache:

$$\widehat{y}_t = \mathrm{Decode}(\mathrm{KV}_t, \textsc{Response}). \tag{27}$$

The content branch therefore inherits the causal clip and bounded action history used for the decision, while avoiding a second visual encoding pass. On NO-PUSH, decoding stops after SILENCE. On PUSH, generation is capped at a short response budget and the recurrent state is committed only

after the model-side generation transaction succeeds. This yields a single-prefill factorization of proactive assistance: the streaming backbone amortizes multimodal perception across intervention and language generation, while TCI reshapes only the action boundary using causal temporal evidence. Consequently, improving intervention selectivity neither invokes a second visual-language forward pass nor replaces the backbone's conditional generation distribution. Negative decisions terminate immediately at the action boundary, whereas positive decisions retain the full BLV-adapted decoder for assistance generation. The bounded recurrent belief and committed-notification ledger make this selective execution independent of the elapsed stream length.

4 EXPERIMENTS

4.1 DATASET

We augment 439 first-person EgoBlind video segments (4.28 hours) (Xiao et al., 2025) with timestamp-aligned supervision for proactive assistance. The videos span seven everyday settings: home activities, indoor public spaces, outdoor environments, transportation, shopping, dining, and travel. Five sighted–BLV annotator dyads followed a complementary protocol: sighted partners provided reliable access to visual evidence, while BLV partners determined whether speech would be useful and authored the assistance message. The resulting dataset contains 7,482 chronological decisions, comprising 1,794 PUSH and 5,688 NO-PUSH annotations.

Complete videos are partitioned by category into 308/67/64 train/validation/test videos, corresponding to 5,313/1,015/1,154 decisions; source-related segments remain within the same partition. Each decision is conditioned on a three-second causal clip sampled at 2 fps with maximum side 384 pixels and on interaction history strictly preceding the target time. The principal comparison uses the 1,090 non-initial test decisions, excluding the mandatory first description in each video so that evaluation focuses on whether newly observed evidence warrants another intervention.

4.2 EXPERIMENTAL SETUP

Compared methods. We compare the pretrained streaming VLM, its BLV-adapted counterpart, and the complete BLV-SceneAgent equipped with TCI. The BLV-adapted baseline combines the SFT-adapted VLM with a separately trained GRU for PUSH/NO-PUSH prediction. TCI is evaluated chronologically under its own previous decisions and committed notifications.

Evaluation metrics. Reference interventions in the evaluation set are annotated with acceptable notification intervals. A predicted PUSH is considered temporally correct if its decision time falls within the corresponding interval. We report macro-F1 and PUSH F1 as primary decision metrics, together with accuracy, PUSH precision, PUSH recall, and average precision (AP). Statistical comparisons use paired bootstrap resampling over complete videos, preserving temporal dependence within each stream. Assistance content uses case-normalized token F1, word-level ROUGE-L (Lin, 2004), and BERTScore-F1 (Zhang et al., 2020). To evaluate the decision–content transaction, each similarity s induces a delivery score

$$\text{D-F1}_s = \frac{2\sum_{t\in\mathcal{T}_{TP}} s(\hat{c}_t, c_t)}{N_{\text{ref}} + N_{\text{pred}}},$$

where $\mathcal{T}_{TP}$ contains correctly triggered reference interventions. The denominator charges both missed assistance and unnecessary speech.

Implementation details. The 8.77B streaming backbone is adapted in BF16 using 174.59M trainable rank-64 rsLoRA parameters, while the vision tower and multimodal projector remain frozen. Training uses two NVIDIA A40 GPUs. TCI adds 476K parameters, including a 96-dimensional recurrent belief, a 64-dimensional intent state, and an eight-entry committed-notification ledger.

4.3 COMPARISON OF METHODS

Table 1 shows that BLV adaptation and temporal intervention contribute at different levels of selective assistance. The pretrained model's 0.255 PUSH precision indicates that many of its proposed

Table 1: Selective intervention performance on the test set.

System	Accuracy	Macro-F1	PUSH P	PUSH R	PUSH F1	AP
Pretrained streaming VLM	0.412	0.408	0.255	0.601	0.358	0.353
BLV-adapted streaming VLM	0.805	0.761	0.631	0.688	0.658	0.664
BLV-SceneAgent (Ours)	**0.861**	**0.824**	**0.754**	**0.732**	**0.743**	**0.756**

Table 2: Leave-one-scene-out transfer of TCI.

Held-out scene	Decisions	PUSH F1	Macro-F1	AP
Dining	126	0.745	0.840	0.715
Home	172	0.729	0.820	0.739
Indoor	122	0.667	0.751	0.705
Outdoor	206	0.703	0.797	0.700
Shopping	119	0.704	0.808	0.658
Transportation	184	0.693	0.788	0.737
Travel	161	0.722	0.791	0.785

Table 3: Ablation study of the temporal mechanisms in TCI.

Variant	PUSH F1	Macro-F1
Full TCI	**0.743**	**0.824**
w/o policy-history refinement	0.686	0.783
w/o repetition-aware suppression	0.710	0.796
w/o change-aware modulation	0.684	0.785

interventions are not supported by the reference decisions. BLV adaptation establishes a stronger decision baseline, while TCI makes the selected reminders more reliable and covers more reference interventions: precision and recall improve together, producing an 8.5-point gain in PUSH F1 over the adapted VLM. The accompanying AP improvement extends this benefit to the ranking of intervention opportunities. With the adapted backbone held fixed, these gains support learning temporal decision structure from scene evolution and committed history as an additional source of assistive capability.

4.4 CROSS-SCENE GENERALIZATION AND ROBUSTNESS

Table 2 examines whether temporal intervention knowledge can be reused across daily activities. In each fold, TCI must apply its learned handling of scene change and previously communicated information to an activity excluded from controller training and validation. The shared BLV-adapted backbone remains fixed, so the transfer is evaluated at the controller level. PUSH F1 ranges from 0.667 to 0.745 across all seven held-out scenes, showing that the learned policy can identify intervention opportunities even without controller supervision from the target activity. This transfer supports using a common temporal intervention design across activities with different visual content and assistance needs.

4.5 ABLATION STUDY

Table 3 highlights two requirements for effective temporal intervention: learning under the policy's own action history and using that history to assess the value of new evidence. Removing policy-history refinement and change-aware modulation produces the largest PUSH F1 drops, at 5.7 and 5.9 points, respectively. These effects support training the controller on histories generated by its own decisions and explicitly incorporating scene evolution into the intervention score. Repetition-aware suppression provides an additional benefit by conditioning that score on previously committed notifications; removing it also reduces both F1 metrics. Together, the ablations support a controller that tracks what has changed alongside what has already been communicated.

4.6 ASSISTANCE CONTENT QUALITY

Table 4 evaluates the messages that each complete system actually elects to deliver. BLV adaptation establishes the largest gain in semantic alignment with BLV-authored assistance. TCI then improves every delivery metric while retaining the identical adapted decoder, isolating a less obvious source of content quality: a message can be useful only when it is delivered at the event it describes. Better temporal selection therefore preserves actionable descriptions and suppresses contextually misplaced speech, coupling the agent's *when-to-speak* and *what-to-say* capabilities.

Table 4: End-to-end assistance-content quality under each system's predicted interventions. Each column is the delivery F1 induced by the corresponding content similarity; higher is better.

System	Tok. D-F1	R-L D-F1	BERT D-F1
Pretrained streaming VLM	0.132	0.121	0.181
BLV-adapted streaming VLM	0.428	0.395	0.526
BLV-SceneAgent (Ours)	**0.483**	**0.446**	**0.594**

Table 5: Model-side latency of the evaluated systems.

Warm model-side action path	P50 (ms)	P95 (ms)
Pretrained streaming VLM	339.8	371.1
BLV-adapted streaming VLM	419.4	445.8
BLV-SceneAgent (full agent)	452.7	478.8
TCI policy transaction	2.4	3.9

Table 6: BLV user ratings of assistance quality.

System	Timeliness ↑	Usefulness ↑	Annoyance ↓	Trust ↑
Pretrained streaming VLM	3.25	2.75	3.40	3.00
BLV-adapted streaming VLM	3.85	3.90	2.85	3.85
BLV-SceneAgent (Ours)	**4.10**	**4.20**	**2.15**	**4.10**

4.7 Efficiency Analysis

All model-side latency measurements are obtained on a single NVIDIA A40 GPU. Table 5 shows that temporal selectivity preserves the streaming model's latency regime: every intervention path remains below 0.5 seconds at P95, including the complete agent. More importantly, the TCI transaction occupies less than 1% of the full P95 path. Its decision gains therefore do not depend on another visual pass or a separate content model. When PUSH is committed, language generation continues from the same multimodal KV cache; when NO-PUSH is selected, decoding terminates immediately. This asymmetric execution concentrates autoregressive cost on the comparatively rare moments at which speech is actually warranted.

4.8 BLV User Study

Six BLV adults (five totally blind and one with low vision) evaluated the pretrained streaming VLM, the BLV-adapted VLM, and BLV-SceneAgent through the same live mobile camera-to-speech interface. Each participant completed two trials per setting with each system in a shopping mall, a restaurant, a living room, and a street. Participants rated timeliness, usefulness, annoyance, and trust on five-point scales; Table 6 reports the mean ratings.

BLV adaptation improves the perceived timeliness and usefulness of assistance, suggesting that relevant content contributes to whether a response is experienced as arriving at the right moment. BLV-SceneAgent further improves these ratings while reducing annoyance and strengthening trust. This pattern suggests that users experience timely assistance through the joint quality of intervention timing and message content. Temporal selectivity helps align what the agent communicates with when that information is useful.

5 Conclusion

We introduced BLV-SceneAgent, a proactive visual assistant that learns when to intervene and what to communicate from BLV-informed chronological supervision. Its TCI policy combines causal scene evolution with committed notification history, while a shared multimodal prefill supports both intervention decisions and content generation. Experiments support temporal selectivity as a source of more useful assistance: information should be novel, actionable, and delivered at the right moment. Future work will explore longitudinal personalization and user-controllable intervention policies.

AI USE STATEMENT

We used AI tools to check the manuscript for ambiguous or erroneous wording and to polish non-standard expressions. We also used AI tools to critically examine our research ideas, helping us identify and discard ideas that were incorrect or insufficiently novel during the research process.

REFERENCES

Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samuel White, and Tom Yeh. VizWiz: Nearly real-time answers to visual questions. In *Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology*, pp. 333–342, 2010. doi: 10.1145/1866029.1866080.

Erin Brady, Meredith Ringel Morris, Yu Zhong, Samuel White, and Jeffrey P. Bigham. Visual challenges in the everyday lives of blind people. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2117–2126, 2013. doi: 10.1145/2470654.2481291.

Ruei-Che Chang, Yuxuan Liu, and Anhong Guo. WorldScribe: Towards context-aware live visual descriptions. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, 2024. doi: 10.1145/3654777.3676375.

Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. VideoLLM-online: Online video large language model for streaming video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18407–18418, 2024.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1724–1734, 2014. doi: 10.3115/v1/D14-1179.

Xin Ding, Hao Wu, Yifan Yang, Shiqi Jiang, Qianxi Zhang, Donglin Bai, Zhibo Chen, and Ting Cao. StreamMind: Unlocking full frame rate streaming video dialogue through event-gated cognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13448–13459, 2025.

Sitong Gong, Tianyu Yan, Caixin Kang, Bo Zheng, Xiang Ruan, Huchuan Lu, Kaipeng Zhang, Yoichi Sato, and Yifei Huang. Vinci2: Providing proactive assistance in continuous egocentric videos. In *European Conference on Computer Vision*, 2026.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. VizWiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3608–3617, 2018.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Junhyeok Kim, Min Soo Kim, Jiwan Chung, Jungbin Cho, Jisoo Kim, Sungwoong Kim, Gyeongbo Sim, and Youngjae Yu. EgoSpeak: Learning when to speak for egocentric conversational agents in the wild. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 2990–3005, 2025. doi: 10.18653/v1/2025.findings-naacl.163.

Walter S. Lasecki, Phyo Thiha, Yu Zhong, Erin Brady, and Jeffrey P. Bigham. Answering visual questions with conversational crowd assistants. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 18:1–18:8, 2013. doi: 10.1145/2513383.2517033.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, 2004.

Daniela Massiceti, Luisa Zintgraf, John Bronskill, Lida Theodorou, Matthew Tobias Harris, Edward Cutrell, Cecily Morrison, Katja Hofmann, and Simone Stumpf. ORBIT: A real-world few-shot dataset for teachable object recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10818–10828, 2021.

Daniela Massiceti, Camilla Longden, Agnieszka Slowik, Samuel Wills, Martin Grayson, and Cecily Morrison. Explaining CLIP's performance disparities on data from blind/low vision users. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12172–12182, 2024.

Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Dispider: Enabling video LLMs with active real-time interaction via disentangled perception, decision, and reaction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24045–24055, 2025.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3982–3992, 2019. doi: 10.18653/v1/D19-1410.

Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635, 2011.

Abigale Stangl, Meredith Ringel Morris, and Danna Gurari. "Person, Shoes, Tree. Is the Person Naked?" what people with vision impairments want in image descriptions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2020. doi: 10.1145/3313831.3376404.

Abigale Stangl, Nitin Verma, Kenneth R. Fleischmann, Meredith Ringel Morris, and Danna Gurari. Going beyond one-size-fits-all image descriptions to satisfy the information wants of people who are blind or have low vision. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 1–15, 2021. doi: 10.1145/3441852.3471233.

Yueqian Wang, Xiaojun Meng, Yuxuan Wang, Jianxin Liang, Jiansheng Wei, Huishuai Zhang, and Dongyan Zhao. VideoLLM knows when to speak: Enhancing time-sensitive video comprehension with video-text duet interaction format. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 6338–6359, 2025. doi: 10.18653/v1/2025.findings-emnlp.336.

Junbin Xiao, Nanxin Huang, Hao Qiu, Zhulin Tao, Xun Yang, Richang Hong, Meng Wang, and Angela Yao. EgoBlind: Towards egocentric visual assistance for the blind. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2025.

Weicai Yan, Yuhong Dai, Qi Ran, Haodong Li, Wang Lin, Tao Jin, Xing Xie, Hao Liao, and Jianxun Lian. Proact-VL: A proactive VideoLLM for real-time AI companions. In *International Conference on Machine Learning*, 2026.

Dingyu Yao, Junhao Zhou, Chenxu Yang, Chuanyu Qin, Haowen Hou, Zheming Liang, Congcong Wang, Yuhang Cao, Shenglong Ye, Shuai Xie, Shuhuan Gu, Haoyang Huang, Qingyi Si, Nan Duan, and Jiaqi Wang. JoyAI-VL-Interaction: Real-time vision-language interaction intelligence. *arXiv preprint arXiv:2606.14777*, 2026.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020.

Yichi Zhang, Xin Luna Dong, Zhaojiang Lin, Andrea Madotto, Anuj Kumar, Babak Damavandi, Joyce Chai, and Seungwhan Moon. Proactive assistant dialogue generation from streaming egocentric videos. *arXiv preprint arXiv:2506.05904*, 2025a.

Ziyi Zhang, Zhen Sun, Zongmin Zhang, Zifan Peng, Yuemeng Zhao, Zichun Wang, Zeren Luo, Ruiting Zuo, and Xinlei He. "I Can See Forever!": Evaluating real-time VideoLLMs for assisting individuals with visual impairments. *arXiv preprint arXiv:2505.04488*, 2025b.