Time-Stratified Architecture for Real-Time Analysis of Behavioural Signals in Telehealth

Dennis Maier[1], Marco Xaver Bornschlegl[1]

[1]FernUniversität in Hagen, Universitätsstraße 1, 58097 Hagen, Germany

Abstract

Telehealth psychotherapy generates heterogeneous behavioural data streams, including facial affect, pose and posture cues, speech transcripts, and linguistic markers, that differ in latency, temporal granularity, and interpretive horizon. Architectures that route these signals through a single processing path risk coupling sub-second in-session support with slower contextual interpretation, while final predictions alone provide limited support for inspection, replay, and later conversational analysis. This paper presents a time-stratified architecture for server-side telehealth analytics that separates processing into three explicit regimes: reactive cue generation (< 250 ms), reflective record materialization (3–10 s emission cadence), and asynchronous long-horizon aggregation and synthesis (30 s+). The central design element is a reflective middle tier that assigns asynchronous multimodal cue streams to event-time windows and materializes them as replayable, window-scoped intermediate records containing cue aggregates, source-context metadata, and modality-availability masks. The architecture operationalizes the AI2VIS4BigData notion of inspectable transformation chains [1] and frames the evaluation in relation to capability, reproducibility, and validity dimensions of the Trustworthy AI-based Big Data Analysis Model (TAI-BDM) [2]. A fully containerized reference implementation is evaluated in five controlled experiments covering concurrent load, frame-level visual modality expansion, synthetic delivery disorder, deterministic replay, and topic-level downstream stress. Results show that the evaluated visual reactive path remains within budget under ten concurrent ENet+Pose+Body sessions (p99 = 110 ms), reflective-window emission coverage remains 100% under the tested synthetic disorder conditions, required record fields remain structurally stable across deterministic replays, and topic-level Tier II stress does not measurably degrade the ENet-only Tier I path (18.8 ± 0.8 ms baseline vs. 18.2 ± 0.4 ms under stress, n=5). These findings support the technical feasibility and reproducibility of the time-stratified architecture.

Keywords

stream processing, multimodal fusion, telehealth, behavioural signals, visual analytics, inspectable intermediate records, AI2VIS4BigData, reproducible deployment

1. Introduction

Telehealth psychotherapy has become an important extension of digital mental-health care [3]. However, video-conference supported sessions can add videoconferencing-related cognitive load for clinicians [4] and may constrain access to non-verbal behaviour. Because psychotherapy is not driven by verbal content alone, clinically relevant interpretation draws on speech- and language-related behaviour [5], facial behaviour [6], body and pose cues [7], and their timing within the dialogue [8]. These modality-specific behavioural phenomena are referred to as *signals*, their machine-derived outputs as *cues*. Such signals and cues have limited value in isolation. A single facial change or cue becomes clinically interpretable only when it is related to interactional context in psychotherapy [9]. This includes conversational structure, such as who was speaking, what was being discussed, and where the signal and cue occurred within the dialogue [8]. Multimodal telehealth analytics therefore generate heterogeneous data at different abstraction levels that do not arrive in a synchronized manner: from raw audio-video streams, the system derives low-level cues within tens of milliseconds, whereas context dependent information emerges only after slower downstream processing. A useful system must support reactive in-session responsiveness, comparable to low-latency serving requirements in online prediction systems [10], across multiple concurrent sessions sharing the same inference infras-

TRI-IVIS 2026: Workshop on Trustable, Reproducible and Intelligent Information Visualization Systems, June 8–12, 2026, Venice, Italy

*Dennis Maier

tructure. It must also organize asynchronous multimodal evidence [11] under event-time semantics for bounded out-of-order arrival [12]. Finally, the resulting derived artifacts should remain inspectable and reusable, consistent with IVIS4BigData [13] and its AI-oriented extension AI2VIS4BigData [1], layered reference models that frame big-data analysis as a transformation chain of inspectable intermediate artifacts spanning data management, analytics, interaction, and visualization. It must also integrate into a videoconferencing setting without overloading end-user devices, and remain reproducible across deployments.

1.1. Research Questions

These constraints imply three architectural requirements: low-latency in-session support, persistent intermediate records that preserve temporally aligned multimodal evidence for inspection, and separation between fast reactive processing, reflective record materialization, and slower non-reactive aggregation and synthesis workloads. Against this background, the paper addresses the following research questions:

RQ1 (Reactive responsiveness): How can a telehealth analytics architecture preserve reactive responsiveness under increasing concurrent session load and richer modality configurations?

RQ2 (Multimodal record materialization): How can heterogeneous multimodal cue streams be assigned to event-time windows and materialized into structurally stable, inspectable intermediate records under bounded out-of-order arrival?

RQ3 (Operational isolation): How can a time-stratified architecture preserve reactive responsiveness while slower contextual processing executes concurrently in a reproducible deployment?

2. State of the Art

The following literature review is structured by the capability requirements implicit in **RQ1**–**RQ3**: reactive responsiveness during live sessions, multimodal record materialization for asynchronous cue streams, and operational isolation between fast and slow processing regimes.

2.1. RQ1: Reactive Responsiveness for Live Telehealth Support

Digital mental-health research has established a broader technological context for mental-health support [3]. Psychotherapy research further frames patient–provider interaction data as a source for structured analytic support [9]. Concrete examples range from projects like SimSensei/Ellie for multimodal assessment [14] to automated therapeutic-dialogue coding [15]. However, these works focus on retrospective coding or narrowly scoped live functionality rather than low-latency multimodal support embedded in ongoing telehealth sessions. General model-serving systems such as Clipper [10] and InferLine [16] address serving-level latency. However, these strands do not yet provide a broader telehealth architecture that preserves reactive support while additional modalities and slower downstream processes execute concurrently, which motivates **RQ1**.

2.2. RQ2: Event-Time Alignment and Inspectable Intermediate Records

Multimodal affective computing provides a general framework for combining heterogeneous modalities [11]. Behavioural signal processing supports cue extraction from speech- and language-related behaviour [5], while facial affect datasets support face-related cues [6] and perception pipelines support body and pose cues [7]. The open problem is not extraction itself but the connection of asynchronous cues to conversational meaning: Most multimodal Machine Learning (ML) work treats temporal alignment as a fusion-internal concern [11], leaving unresolved how heterogeneous outputs should be synchronized, stabilized, and materialized as reusable intermediate records. Therapeutic conversation is sequentially organized as interactional practice [8]. More generally, conversational coordination

depends on turn-taking structures [17] and fine-grained pauses, gaps, and overlaps [18]. A cue misaligned in time may therefore be attached to the wrong speaker, utterance, or local sequence. Research in healthcare AI cautions that final predictions alone are not sufficient for clinical use [19]. For use in clinical environments, transparency and interpretability are likewise emphasized as conditions for understandability [20]. IVIS4BigData [13] and AI2VIS4BigData [1] frame this need as a transformation chain in which derived artifacts remain inspectable. What remains open for **RQ2** is whether multimodal cues can be assigned to event-time windows and materialized into stable intermediate records that remain usable under bounded out-of-order arrival and can serve as a substrate for later conversational mapping.

2.3. RQ3: Operational Isolation of Fast and Slow Processing Regimes

General-purpose stream-processing architectures provide relevant primitives, such as Lambda [21], Kappa [22], and the Dataflow Model's event-time semantics [12], but remain domain-agnostic. At the serving level, Clipper [10] and InferLine [16] address latency-aware online prediction. Microservice-oriented designs address service decomposition and operational modularity [23]. However, telehealth analytics requires these capabilities to operate simultaneously: reactive inference must remain responsive during the session while slower contextual aggregation, persistence, and long-horizon synthesis proceed in parallel. The unresolved issue behind **RQ3** is therefore not whether individual primitives exist, but whether they can be assembled into one reproducible telehealth architecture in which slower non-reactive workloads do not interfere with the reactive path.

2.4. Remaining Challenges

The reviewed literature provides partial capabilities but leaves their integration into a solid, fast, multimodal grounded, inspectable, and deployable telehealth architecture unresolved. This uncovers three remaining challenges:

RC1 **Reactive in-session support remains insufficiently evidenced under concurrent telehealth session load** (**RQ1**): existing systems show that live inference and scalable serving are possible, but provide limited evidence that a telehealth analytics architecture can preserve reactive responsiveness as concurrent session load increases.

RC2 **Multimodal cues are insufficiently materialized into stable intermediate records as a prerequisite for later conversational mapping** (**RQ2**): prior work does not resolve how asynchronous cue streams should be assigned to event-time windows and preserved as intermediate records that can later be related to conversational units under bounded out-of-order arrival.

RC3 **Operational isolation of fast and slow regimes is insufficiently demonstrated** (**RQ3**): limited evidence that reactive and contextual processing can coexist in one reproducible deployment without degrading the reactive path.

These three challenges jointly motivate the time-stratified artifact presented next.

3. Artifact Design and Reference Implementation

This work contributes a time-stratified architecture for multimodal telehealth analytics. The resulting material is a layered processing model in which raw channels, derived cue streams, intermediate records, and later conversational artifacts remain distinguishable in both time and representation and consistent with the layered perspective of AI2VIS4BigData [1].

3.1. Time-Stratified Processing Model

The design principle is a decomposition into three latency regimes (Table 1). These regimes separate reactive support (RC1), reflective record materialization (RC2), and non-reactive long-horizon aggregation and synthesis (RC3), with explicit operational boundaries between them. Tier I is reactive and

Table 1
Time-stratified processing regimes.

Tier	Budget	Primary product	Primary role	Main consumer
I	< 250 ms	Cue events	Reactive in-session support	Live awareness
II	3–10 s cadence	Intermediate records	Short-horizon record materialization	Review and contextual inspection
III	30 s+	Retrospective and synthesized artifacts	Long-horizon aggregation, agentic synthesis, reporting	Clinical artifacts and downstream review

produces low-latency modality-specific cue events for direct session support; Tier II is reflective and assigns asynchronous Tier I outputs to bounded short-horizon event-time windows that are materialized as inspectable intermediate records; Tier III is asynchronous and non-reactive. It consumes persisted Tier I and Tier II artifacts over longer temporal horizons to derive retrospective aggregates, conversational summaries, algorithmic patterns, and agent- or Large Language Model (LLM)-based clinical artifacts. These workloads may run periodically during an ongoing session or after session closure; they are not benchmarked in this workshop paper and do not participate in the reactive path. The < 250 ms Tier I budget is used as an operational design budget for the direct reactive path. This design choice is motivated by the fine temporal granularity of turn-taking [17], pauses, gaps, and overlaps in conversation [18], and general expectations for low-latency interactive communication [24]. It is therefore not used as a clinical adequacy threshold, but as the boundary that separates immediate Tier I responsiveness from the larger reflective and asynchronous time ranges of Tier II and Tier III. Because derived signals from raw audio/video fan out at different granularities and delays, a single uniform processing clock is inadequate (see Figure 1). For Tier II, the implementation uses a short sliding-window configuration (Sec. 3.3) to accumulate delayed multimodal evidence for RC2 while keeping reflective updates within the 3–10 s inspection cadence in Table 1.

3.2. Intermediate Records and Conversational Mapping

Tier II emits window-scoped intermediate records that capture event-time-windowed multimodal cue summaries, temporal boundaries, modality-availability information, and lightweight source references (**RC2**). The records are intended as a visualization substrate: they expose temporal scope, modality availability, cue aggregates, and source context for later inspection rather than hiding them inside final predictions. Unlike Tier I, Tier II is not designed to provide immediate responses; it aggregates longer local time intervals, allowing affective and behavioural cues to be examined in conjunction with additional temporal and multimodal contextual information. It is therefore calculated using limited sliding windows or custom ranges, which makes this layer the decoupled main analysis layer. Tier III can use persisted records for long-horizon aggregation and synthesis by relating windows, turns, topics, entities, segments, and clinical artifacts through explicit temporal overlap rules. This concept separates the *computation clock* (window-based) from the *interpretation clock* (conversation-linked), yielding a stable reflective layer (Figure 1). Inspectability is thus a structural property of the materialized records; the user-facing inspection and visualization interface itself is out of scope for this workshop paper and is left to future work.

3.3. Reference Architecture and Tier Services

Ingress is handled by a LiveKit [25] Selective Forwarding Unit (SFU) over WebRTC [26]. Tier I inference runs in a Ray-based serving layer [27] with fractional GPU sharing and asynchronous actor concurrency. The implementation hosts an EfficientNet (ENet)-based facial-affect actor [28] trained on AffectNet [6], a MediaPipe pose actor [7], a pose-derived body-signal actor, a Whisper Automatic Speech Recognition (ASR) actor [29], and a CPU-only spaCy/scispaCy-based Named Entity Recognition (NER) actor [30]. These services emit asynchronous cue streams rather than an immediately fused result, pre-

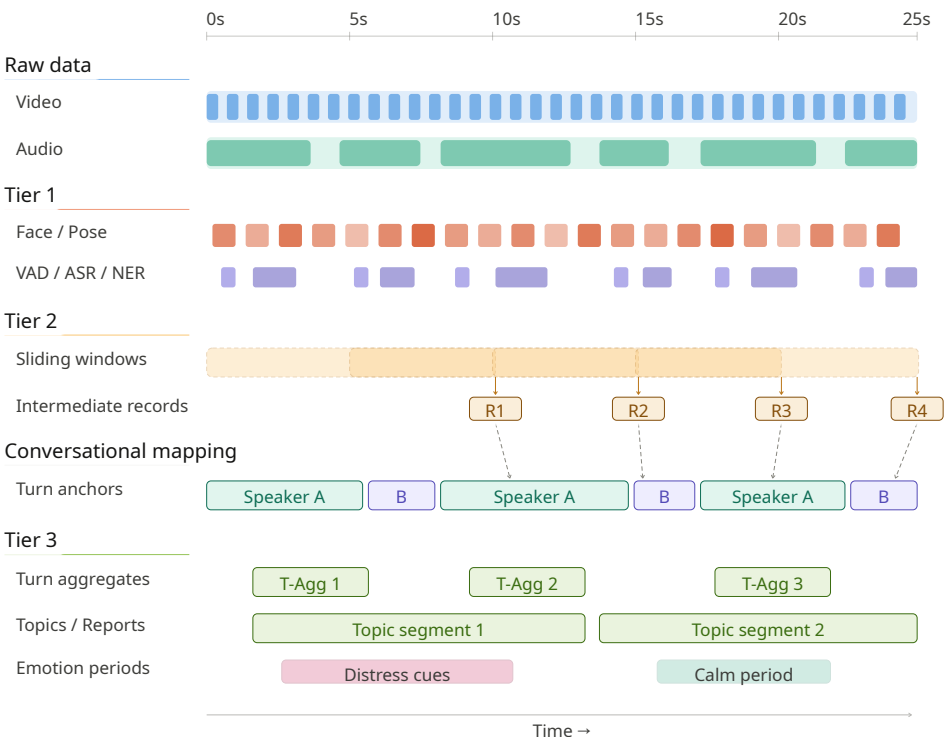


Figure 1: Temporal stratification across the full model. Fast cue streams are assigned to bounded windows and related over longer horizons to conversational and retrospective artifacts. Tier III denotes non-reactive aggregation and synthesis workloads that are part of the model but not benchmarked in this workshop paper.

serving the reactive path while enabling later reflective record materialization. Transcript-dependent entity extraction is invoked asynchronously. Kafka [31] decouples inter-tier transport. Tier II record materialization uses PyFlink for stateful stream processing [32]. In the benchmark configuration, Tier II uses 30 s sliding windows with a 10 s slide to accumulate delayed multimodal evidence while still emitting near-live reflective updates. Tier III is modeled as a non-reactive long-horizon layer that consumes persisted Tier I and Tier II outputs to derive turn-level and topic-linked aggregates, emotion periods, summaries, algorithmic patterns, and LLM-based session synthesis. Inference and reflective record materialization execute server-side; end-user devices remain limited to videoconferencing and frontend rendering. These window parameters and model choices are representative engineering settings rather than clinically optimal or universally optimal configurations; the architectural claims are orthogonal to the specific cue extractors used.

3.4. Event-Time Materialization, Persistence, and Orchestration

Reflective materialization uses event-time semantics and bounded-out-of-orderness watermarks [12]. Here, *event time* denotes the timestamp at which a cue was produced at the source, as opposed to the processing time at which it reaches Tier II; a *watermark* is an advancing bound asserting that no further events with a timestamp below it are expected, which is what allows a window to be finalized under out-of-order arrival. In the reference implementation, the watermark bound is configured to 5 s, i.e. events arriving up to 5 s late are still assigned to their event-time window. The reflective layer aggregates cue trajectories, tracks modality availability, enriches records when transcript-dependent cues arrive, and materializes a window-scoped intermediate record. A modality is marked unavailable when no valid event falls into the window after event-time assignment. For **RC3**, emitted records retain source-context metadata, source hashes, and evidence references as they move through Kafka, so downstream artifacts can be traced back to predecessor records. Tier II records are persisted as JSON in PostgreSQL, while high-volume Tier I evidence and longer-lived archives are written to MinIO as Parquet. Decoupled transport makes regime boundaries benchmarkable independently.

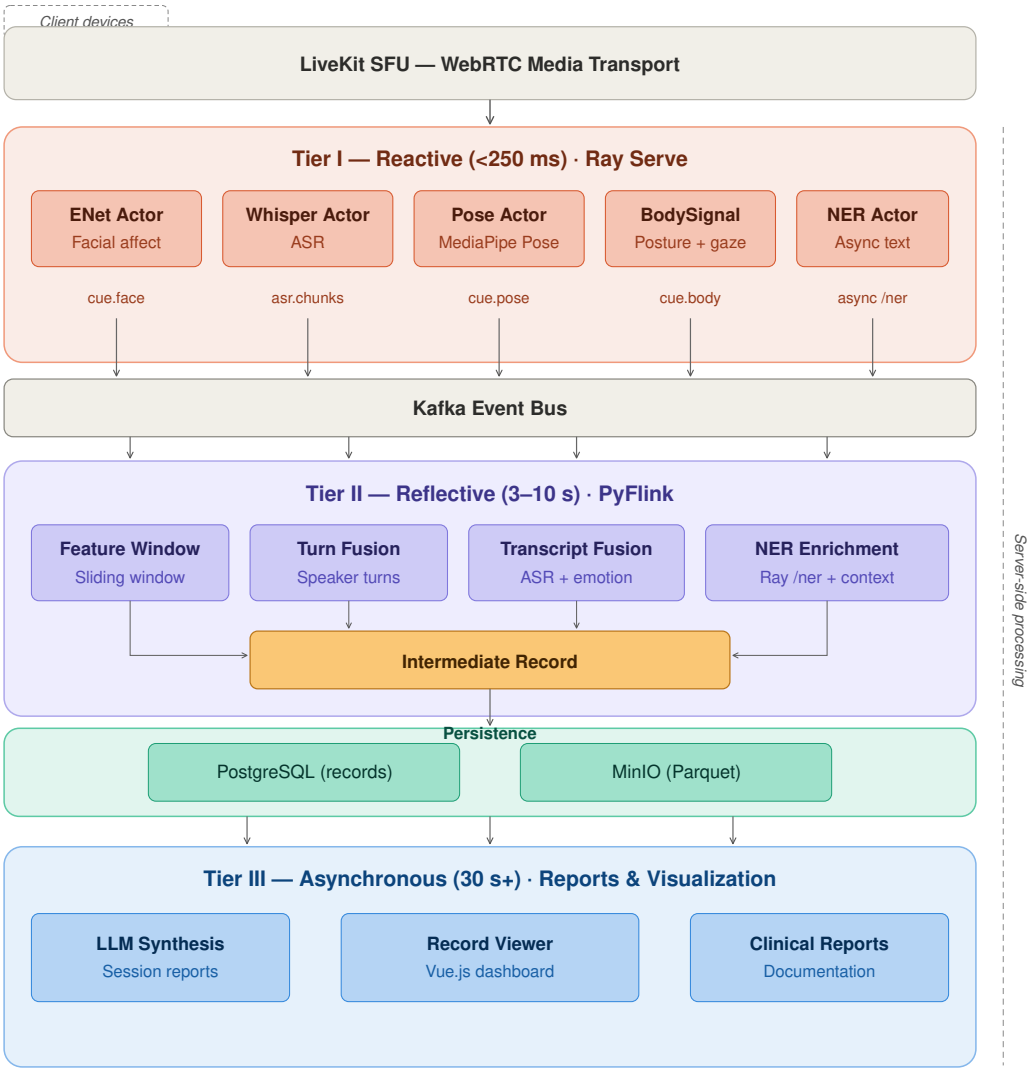


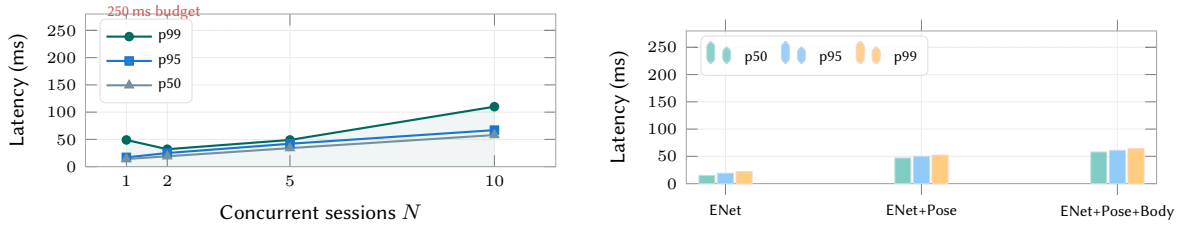
Figure 2: Reference architecture. Raw channels are transformed into asynchronous cue streams, assigned to reflective windows and materialized as intermediate records, and reused by non-reactive long-horizon aggregation and synthesis workloads.

4. Experimentation and Evaluation

This section evaluates the reference implementation across RC1 to RC3. As a workshop evaluation, it focuses on reactive latency, reflective record emission, structural consistency, and downstream-load isolation under controlled workloads. Clinical validity, semantic conversational alignment, user-facing visualization, workflow quality, and Tier III aggregation and synthesis workloads are reserved for follow-up work. All experiments use deterministic, seed-controlled workload generation with fixed workload parameters and pinned container versions on a single workstation (Intel Core Ultra 9 285K, NVIDIA RTX 5090, 96 GB RAM). This evaluation maps the RCs to five experiments: E1 and E2 address **RC1**/**RQ1** by measuring Tier I latency under concurrent load and expanded frame-level modality configuration; E3 and E4 address **RC2**/**RQ2** by measuring reflective session-level window emission and structural stability under controlled disorder and deterministic replay; E5 addresses **RC3**/**RQ3** by testing whether downstream reflective load propagates back into the reactive path.

4.1. RC1: Reactive Responsiveness under Load

E1 Concurrency sweep: Frame-level visual-path latency in Tier I was measured for $N \in \{1, 2, 5, 10\}$ concurrent sessions, each running the ⟨ENet+Pose+Body⟩ visual path (Body denotes the pose-derived



(a) Concurrency sweep.

(b) Modality expansion.

Figure 3: Tier I frame-level visual-path latency percentiles. Left: vs. concurrent sessions; right: by frame-level modality configuration. p99 stays below the 250 ms budget throughout.

body-signal actor; 10 s warmup, 30 s measurement). Tier I stayed within the 250 ms 99th-percentile (p99) latency budget throughout (p99 = 110 ms, p95 = 67 ms, p50 = 58 ms at $N{=}10$) (Figure 3 left). **E2 Modality expansion:** Three configurations were compared over 1500 frames each (3 runs, median reported, warmup per configuration), driven by a single person frame so that all analyzers run under full active inference: (A) ENet, (B) ENet+Pose, (C) ENet+Pose+Body, where Body denotes the body-signal actor deriving posture, head-pose, and gaze features. p99 rose from 22 ms (ENet) to 52 ms (ENet+Pose) and 64 ms (ENet+Pose+Body) (Figure 3 right); each added modality contributes a measurable but bounded tail-latency increment. Whisper and NER run asynchronously and are not part of this GPU-path measurement. The measured visual reactive path remains well inside the Tier I budget, at roughly one quarter of the 250 ms bound.

4.2. RC2: Reflective Record Emission under Disorder

E3 and **E4** test if the reflective layer assigns asynchronous cues to event-time windows and materializes them into stable Tier II intermediate records. The conversational-mapping correctness itself is a Tier-III concern and out of scope for this workshop. **E3** and **E4** target coverage and structural consistency as the necessary conditions for later mapping. These metrics evaluate reflective emission and schema-level stability, not semantic correctness of conversational alignment. **E3 Emission under disorder:** Three concurrent sessions emitted 200 events each (≈20 s per session at 10 Hz) while out-of-order rates were swept over $\{0, 5, 10, 20\}$%, implemented as $\pm$3-position shuffles ($\approx \pm$300 ms at 10 Hz). Coverage is measured at session granularity: the fraction of sessions whose window is finalized (on watermark advance, including at end of stream) into $\geq$1 Tier II record, so the three sessions yield three expected records. It remained 100% across all disorder levels. As the injected disorder ($\pm$300 ms) is small relative to the 5 s watermark and 30 s window, this is a coverage and event-time-assignment check under mild reordering rather than a late-data stress test. The workload is fully synthetic and serves as a controlled baseline rather than a field-data distribution. **E4 Repeated-run structural stability:** A fixed synthetic batch (2 sessions $\times$ 60 events) was replayed three times under deterministic seeds and pinned containers. Across all replays, the session coverage remained 100%. The tested fields `face` and `fused` were consistently present, indicating structural field consistency under controlled repetition.

4.3. RC3: Isolation of the Reactive Path

E5 Downstream stress isolation: A three-phase experiment measured Tier I ⟨visual-path / ENet-only⟩ latency during 180 s baseline, 240 s downstream stress, and 180 s recovery (10 min per run, $n{=}5$). During stress, synthetic events were injected directly into the Tier II Kafka topic at $5\times$ the normal rate (≈59,700 events per phase), bypassing Tier I. As this injected load is transport- and CPU-level work on Tier II while the measured path is GPU-bound ENet inference, E5 tests isolation at the Kafka topic boundary rather than full shared-resource contention. Across five runs, Tier I p99 measured 18.8 ± 0.8 ms at baseline and 18.2 ± 0.4 ms under computational load (mean change $-3.0 \pm 6.0\%$); recovery returned to 18.8 ± 0.8 ms. Figure 4 shows a representative run and no sustained phase-level

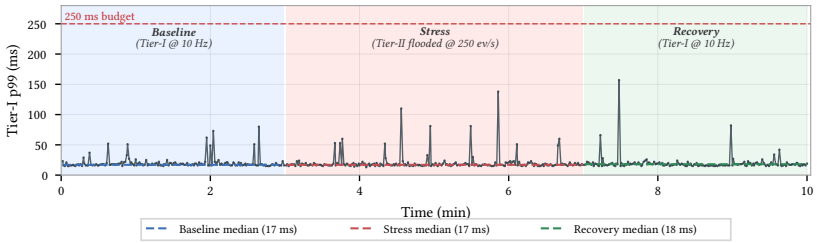


Figure 4: Tier I p99 latency during the downstream stress experiment evaluating RC3. Background shading marks the three phases; dashed lines show per-phase medians.

increase is visible. In summary, under this topic-level stress configuration, E5 provides evidence that additional Tier II load did not measurably degrade the Tier I reactive latency path.

5. Discussion

Taken together, E1 to E5 support the architectural claim that reactive support, reflective record materialization, and slower long-horizon aggregation and synthesis can be separated without collapsing into a single monolithic path. In E1 and E2, Tier I remained within its latency limits under concurrency and modality expansion (**RQ1**). The modality-expansion results should be read as budget-compliance evidence: under full active inference each added frame-level modality raised Tier I p99 by a measurable but bounded increment, remaining at roughly one quarter of the 250 ms budget. In E3 and E4, the reflective layer maintained stable record emission and structural consistency under controlled disorder and replay (**RQ2**). E5 showed that downstream reflective stress did not measurably degrade the reactive path in five repeated runs under the configured topic-level stress condition (**RQ3**).

In conclusion, these results should therefore be read as evidence for technical feasibility of the time-stratified architecture proposed above, not as evidence for clinical adequacy. The RC3 experiment demonstrates isolation at the configured transport and topic boundary; it does not yet cover full node-level resource contention, broker saturation, storage back pressure, or production-scale distributed deployment. Beyond these measured properties, reducing clinician cognitive load in video-mediated sessions [4] is a downstream goal that the architecture enables but does not itself measure. Temporal scope is preserved and modality context is explicitly represented, but clinical appropriateness, semantic completeness, and conversational-mapping correctness remain unevaluated.

All experiments use synthetic workloads; no patient data is used. In a real-world implementation, the following would be required prior to clinical use: informed consent, data minimisation, encryption, data retention and audit policies, supervision by healthcare professionals, and domain validation. Limitations bound these claims to a single-workstation deployment, synthetic workloads, $N{=}10$ concurrency, ± 300 ms shuffles, and an architectural model in which Tier III aggregation and synthesis workloads are not benchmarked; future work targets these long-horizon workloads, conversational-mapping ground truth, and distributed deployments.

6. Conclusion

This workshop paper presented a time-stratified architecture for multimodal behavioural-signal analysis in telehealth psychotherapy video conferences. The process is divided into three distinct areas: reactive cue generation, reflective materialization of inspectable intermediate records, and asynchronous long-horizon aggregation and synthesis, connected through decoupled, non-blocking transport and persistence boundaries. The contribution is an architectural design artifact: a time-layered design that

treats verifiable interim results as outputs and provides an empirical and technical foundation for subsequent multimodal, dialogue-oriented validation. Within the tested reference implementation, the results demonstrate Tier I responsiveness under the evaluated concurrency and frame-level modality configurations, reflective record materialization under controlled disorder and replay, and cross-regime isolation under topic-level downstream stress. Tier III workloads remain part of the architecture but, being non-reactive long-horizon tasks rather than time-critical reactive ones, are not evaluated in this workshop paper.

7. Declaration on Generative AI

In the process of writing this work, the authors used Anthropic's Claude for grammar and spelling checks and for suggestions concerning word choice and writing style. The academic content, data, analyses, methods, results and conclusions are the work of the authors themselves. The references were selected by the authors. All suggestions to improve readability were reviewed by the authors. The authors take full responsibility for the content of the publication.

References

[1] T. Reis, M. X. Bornschlegl, M. L. Hemmje, AI2VIS4BigData: A reference model for ai-based big data analysis and visualization, in: Advanced Visual Interfaces. Supporting Artificial Intelligence and Big Data Applications, volume 12585 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 1–18.

[2] M. X. Bornschlegl, Towards Trustworthiness in AI-based Big Data Analysis, Ph.D. thesis, FernUniversität in Hagen, 2024. doi:10.18445/20240426-221234-0.

[3] J. Torous, S. Bucci, I. H. Bell, et al., The growing field of digital psychiatry: Current evidence and the future of apps, social media, chatbots, and virtual reality, World Psychiatry 20 (2021) 318–335.

[4] R. Riedl, On the stress potential of videoconferencing: Definition and root causes of zoom fatigue, Electronic Markets 32 (2022) 153–177.

[5] S. Narayanan, P. G. Georgiou, Behavioral signal processing: Deriving human behavioral informatics from speech and language, Proceedings of the IEEE 101 (2013) 1203–1233.

[6] A. Mollahosseini, B. Hasani, M. H. Mahoor, AffectNet: A database for facial expression, valence, and arousal computing in the wild, IEEE Transactions on Affective Computing 10 (2019) 18–31.

[7] C. Lugaresi, J. Tang, H. Nash, et al., MediaPipe: A framework for building perception pipelines, arXiv preprint arXiv:1906.08172, 2019.

[8] A. Peräkylä, C. Antaki, S. Vehviläinen, I. Leudar, Conversation Analysis and Psychotherapy, Cambridge University Press, 2008.

[9] Z. E. Imel, M. Steyvers, D. C. Atkins, Computational psychotherapy research: Scaling up the evaluation of patient–provider interactions, Psychotherapy 52 (2015) 19–30.

[10] D. Crankshaw, X. Wang, G. Zhou, et al., Clipper: A low-latency online prediction serving system, in: Proceedings of NSDI, 2017, pp. 613–627.

[11] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (2019) 423–443.

[12] T. Akidau, R. Bradshaw, C. Chambers, et al., The dataflow model: A practical approach to balancing correctness, latency, and cost in massive-scale, unbounded, out-of-order data processing, Proceedings of the VLDB Endowment 8 (2015) 1792–1803.

[13] M. X. Bornschlegl, K. Berwind, M. Kaufmann, et al., IVIS4BigData: A reference model for advanced visual interfaces supporting big data analysis in virtual research environments, In *AVI Workshop on Big Data Applications*. Springer, 2016.

[14] D. DeVault, R. Artstein, G. Benn, et al., SimSensei kiosk: A virtual human interviewer for healthcare decision support, in: Proceedings of AAMAS, 2014, pp. 1061–1068.

[15] M. Tanana, K. A. Hallgren, Z. E. Imel, et al., A comparison of natural language processing methods for automated coding of motivational interviewing, Journal of Substance Abuse Treatment 65 (2016) 43–50.

[16] D. Crankshaw, G. Sela, X. Mo, et al., InferLine: Latency-aware provisioning and scaling for prediction serving pipelines, in: Proceedings of SoCC, 2020, pp. 477–491.

[17] T. Stivers, N. J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, G. Hoymann, F. Rossano, J. P. de Ruiter, K.-E. Yoon, S. C. Levinson, Universals and cultural variation in turn-taking in conversation, Proceedings of the National Academy of Sciences 106 (2009) 10587–10592. doi:10.1073/pnas.0903616106.

[18] M. Heldner, J. Edlund, Pauses, gaps and overlaps in conversations, Journal of Phonetics 38 (2010) 555–568. doi:10.1016/j.wocn.2010.08.002.

[19] M. Ghassemi, L. Oakden-Rayner, A. L. Beam, The false hope of current approaches to explainable artificial intelligence in health care, The Lancet Digital Health 3 (2021) e745–e750.

[20] D. W. Joyce, A. Kormilitzin, K. A. Smith, A. Cipriani, Explainable artificial intelligence for mental health through transparency and interpretability for understandability, npj Digital Medicine 6 (2023) 6.

[21] N. Marz, J. Warren, Big Data: Principles and Best Practices of Scalable Real-Time Data Systems, Manning Publications, 2015.

[22] J. Kreps, Questioning the lambda architecture, O'Reilly Radar, 2014.

[23] N. Dragoni, S. Giallorenzo, A. L. Lafuente, et al., Microservices: Yesterday, today, and tomorrow, in: Present and Ulterior Software Engineering, Springer, 2017, pp. 195–216.
[24] ITU-T, Recommendation G.114: One-way transmission time, Technical Report, International Telecommunication Union, 2003.
[25] LiveKit Inc., LiveKit: Open source webrtc infrastructure, 2024. URL: https://livekit.io.
[26] A. Bergkvist, D. C. Burnett, C. Jennings, A. Narayanan, WebRTC: Real-Time Communication in Browsers, Recommendation, W3C, 2021.
[27] P. Moritz, R. Nishihara, S. Wang, et al., Ray: A distributed framework for emerging ai applications, Proceedings of OSDI, 2018.
[28] M. Tan, Q. V. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, in: Proceedings of ICML, 2019, pp. 6105–6114.
[29] A. Radford, J. W. Kim, T. Xu, et al., Robust speech recognition via large-scale weak supervision, in: Proceedings of ICML, 2023, pp. 28492–28518.
[30] M. Neumann, D. King, I. Beltagy, W. Amsel, ScispaCy: Fast and robust models for biomedical natural language processing, in: Proceedings of the BioNLP Workshop, 2019, pp. 319–327.
[31] Apache Software Foundation, Apache Kafka: A distributed streaming platform, 2024. URL: https://kafka.apache.org.
[32] Apache Software Foundation, Apache Flink: Stateful computations over data streams, 2024. URL: https://flink.apache.org.