한국컴퓨터그래픽스학회
Korea Computer Graphics Society Vol. 32, No. 3, P. 185~195

AI VTuber 체감 반응성 향상을 위한 지연 은폐 아키텍처

이영성† 최재효 김주원 김세하 한다성*

한동대학교

{22573001, pikpie, 22673001, 22301003}@handong.ac.kr, dshan@handong.edu

A Latency-Cover Architecture for Enhancing Perceived Responsiveness in AI VTubers

Youngsung Ree† Jaehyo Choi Joowon Kim Seha Kim Daseong Han*

Handong Global University

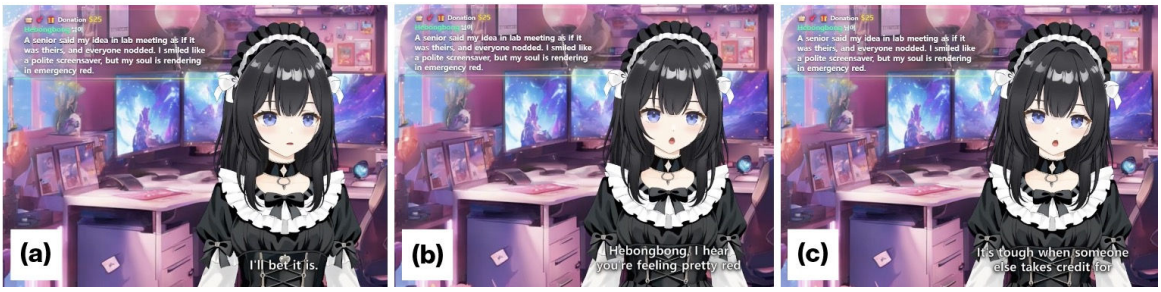

Figure 1: Example responses generated by our study: (a) FastTrack reaction selected using emotion evidence only, (b) grounded FastTrack reaction selected using both emotion and dialogue-act evidence, and (c) the subsequent SlowTrack response providing the full conversational content.

요 약

LLM 기반 AI VTuber의 언어 생성, 음성 합성, 아바타 표현에서 발생하는 누적 지연은 실시간 방송 중 눈에 띄는 침묵을 유발할 수 있다. 본 연구는 이를 시청자의 체감 반응성 관점에서 재정의하고, 정서와 대화행위에 근거한 사전 음성화 선행 반응을 즉시 출력하는 FastTrack과 맥락 기반 본문을 생성하는 SlowTrack으로 구성된 이중 트랙 지연 은폐 아키텍처를 제안한다. 32명을 대상으로 한 반복측정 평가에서 캐릭터 매력도는 조건 간 유의한 차이를 보였으며(Friedman $p = .009$), Emotion Only 조건은 SlowTrack Only 조건보다 높은 평가를 받았다(Holm 보정 $p = .043$). 다만 대부분의 다른 지표에서는 유의한 조건 차이가 확인되지 않았다. 본 연구는 AI VTuber의 지연 문제를 연산 최적화에서 지각적 상호작용 설계로 확장하고, 선행 반응의 맥락 적합성을 평가하는 프레임워크를 제시한다.

Abstract

Accumulated latency from language generation, speech synthesis, and avatar rendering can introduce noticeable silence in live AI VTuber broadcasts. This study reframes this problem in terms of perceived responsiveness and proposes a dual-track latency-cover architecture. FastTrack immediately plays a pre-synthesized preliminary reaction grounded in user emotion and dialogue act, while SlowTrack generates the context-aware main response. In a repeated-measures evaluation with 32 participants, Character Appeal differed significantly across conditions (Friedman $p = .009$), and the Emotion Only condition was rated higher than the SlowTrack Only condition (Holm-adjusted $p = .043$); however, most other measures showed no significant condition effects. The proposed architecture shifts the focus from computational optimization alone to perceptual interaction design and provides a framework for evaluating the contextual appropriateness of preliminary reactions.

키워드: 인공지능 VTuber, 대규모 언어 모델, 체감 반응성, 사회적 존재감, 체화형 대화 에이전트
Keywords: AI VTuber, Large Language Model, Perceived Responsiveness, Social Presence, Embodied Conversational Agent

*corresponding author: Daseong Han / Handong Global University (dshan@handong.edu)

Received : 2026.05.29./ Review completed : 1st 2026.06.07. 2nd 2026.06.25./ Accepted : 2026.06.29.
DOI : 10.15701/kcgs.2026.32.3.185
ISSN : 1975-7883(Print)/2383-529X(Online)

1 서론

라이브 스트리밍은 시청자가 단순히 완성된 영상을 소비하는 방식이 아니라, 실시간 채팅, 후원 메시지, 이모지, 밈(Meme), 시청자 커뮤니티를 통해 방송자와 즉각적으로 상호작용하는 미디어 형식으로 자리 잡았다. 이러한 환경에서 Virtual YouTuber (VTuber)는 2D 또는 3D 가상 아바타를 통해 실시간 방송을 수행하는 새로운 형태의 디지털 퍼포머로 등장하였다 [1]. VTuber 방송에서 시청자는 화면에 보이는 캐릭터를 단순한 그래픽 객체로 받아들이지 않는다. 오히려 아바타의 외형, 목소리, 페르소나, 말투, 실시간 채팅 반응, 방송 중 형성되는 서사를 종합적으로 해석하며, 이를 통해 가상 캐릭터와 방송자 사이의 복합적 정체성을 경험한다 [1, 2].

기존 실사 스트리밍에서 시청자 참여는 방송자의 실제성, 즉흥성, 실시간 반응성에 의해 강화되는 경우가 많았다. 반면 VTuber 환경에서는 실제 수행자의 모습이 직접 노출되지 않고, 아바타가 방송자의 사회적 얼굴로 기능한다. 이로 인해 시청자는 가상성과 실제성이 혼합된 방식으로 방송자를 해석하며, 캐릭터의 일관성, 가상 세계관, 연기된 페르소나, 실제 인간 수행자의 흔적을 동시에 고려한다 [2, 3]. 최근 avatarized livestreamer 연구 역시 VTuber 시청 경험이 단순한 익명성이나 캐릭터 연기가 아니라, 친밀감, 진정성, 극적 몰입, 캐릭터 지속성, 실제 수행자와 가상 캐릭터의 관계 해석을 포함하는 복합적 상호작용임을 보여준다 [3].

최근에는 인간 연기자가 아바타를 조작하는 전통적 VTuber를 넘어, Artificial Intelligence (AI) 기반 에이전트가 방송자 또는 공동 진행자로 참여하는 AI streamer에 대한 관심도 증가하고 있다 [4]. Large Language Model (LLM)의 발전으로 AI 에이전트는 보다 자연스럽고 개방적인 대화를 수행할 수 있게 되었으며, 캐릭터 페르소나를 유지하면서 시청자와 상호작용하는 AI VTuber의 구현 가능성도 높아지고 있다. 그러나 AI streamer는 일반적인 일대일 챗봇과 달리, 다수의 시청자가 동시에 관찰하고 평가하는 공개적이고 공연적인 live streaming 환경에서 작동한다. 따라서 이 환경에서 에이전트는 한 명의 사용자에게 적절한 답을 생성하는 것뿐 아니라, 채팅 흐름, 방송의 리듬, 캐릭터 페르소나, 시청자와의 관계, 아바타의 표현, 후원 메시지에 대한 반응까지 함께 고려해야 한다 [4].

이 과정에서 응답 지연(latency)은 AI VTuber가 직면하는 핵심 문제 중 하나이다. AI VTuber는 시청자의 입력을 해석한 뒤, LLM 기반 응답 생성, Text-to-Speech (TTS) 합성, 아바타 표현, lip-sync 등의 과정을 거쳐 최종 반응을 제공하며, 이러한 다중양식 처리 과정은 누적 지연을 유발할 수 있다 [5, 6]. 라이브 방송 환경에서는 그 지연이 단순한 처리 시간이 아니라 시청자가 체감하는 침묵 구간으로 나타난다. 특히 체화형 대화 에이전트 연구에서는 응답을 기다리는 동안 제공되는 짧은 언어적 · 비언어적 피드백이 체감 지연과 몰입에 영향을 줄 수 있음이 보고되었다 [7]. 따라서 실시간 방송에서의 무응답 구간은 캐릭터가 멈추었거나 시청자의 발화를 놓쳤다는 인상을 줄 수 있는 상호작용 문제로 다룰 필요가 있다.

본 연구는 이러한 관점에서 AI VTuber의 응답 지연을 체감 반응성(perceived responsiveness)의 문제로 재해석하고, 짧은 선행 반응을 통해 대기 시간을 완화하는 latency-cover 접근을 제안한다. 기존 접근은 주로 더 좋은 LLM, 더 자연스러운 TTS, 더 정교한 아바타 표현, 또는 전체 시스템 지연을 줄이는 최적화에 초점을 두어 왔다. 그러나 방송형 AI VTuber에서는 장문의 응답을 생성하는 동안 발생하는 짧은 침묵도 시청자에게 상호작용 단절로 지각될 수 있다. 또한 빠른 반응이 항상 좋은 경험을 보장하는 것도 아니다. 입력 발화의 정서나 대화 맥락과 맞지 않는 짧은 반응은 무작위 filler처럼 보일 수 있고, 사용자는 이를 시스템이 지연을 감추기 위해 삽입한 인위적 발화로 받아들일 수 있다.

본 연구는 다음 질문으로 출발한다. “LLM 기반 장문의 응답을 기다리는 동안, 데이터셋 근거에 기반하여 사전 음성화된 짧은 반응을 먼저 출력하면 AI VTuber의 체감 응답성과 방송 연속성을 개선할 수 있는가?” 이 질문에서 중요한 점은 짧은 선행 반응을 완결된 답변 생성 경로로 보지 않는다는 것이다. 본 연구에서 선행 반응은 사용자의 질문에 대한 최종 답을 제공하는 발화가 아니라, 장문 응답이 준비되는 동안 방송 공백을 덜 어색하게 만들고 시청자가 캐릭터의 반응성을 지각하도록 돕는 cover reaction으로 정의된다. Figure 1는 이러한 선행 반응과 후속 본문 응답의 예를 보여준다.

이를 위해 본 논문은 Dual-track 아키텍처를 제안한다. 이 아키텍처는 AI VTuber의 응답을 하나의 생성 경로로 처리하지 않고, 즉각적인 짧은 반응을 담당하는 FastTrack과 장문 본문 응답을 담당하는 SlowTrack으로 분리하는 latency-cover 아키텍처이다. FastTrack은 사용자의 발화를 정서와 대화행위 관점에서 분석하고, 데이터셋 근거에 기반한 짧은 반응을 사전 음성화된 형태로 제공한다. SlowTrack은 더 긴 문맥, 캐릭터 페르소나, 방송 맥락을 반영하여 본문 응답을 생성한다. 본 논문의 기술적 기여는 새로운 감정 분류 모델이나 새로운 dialogue-act 모델 자체를 제안하는 데 있지 않다. 오히려 검증된 자연어 처리(NLP) 데이터셋과 분류 모듈, 사전 음성화 반응 풀, 방송형 VTuber 런타임, 장문 LLM 응답 경로를 시간적으로 분리 · 조정하여 체감 지연을 줄이는 시스템 아키텍처와 평가 관점을 제시하는 데 있다. 본 논문의 기여는 다음과 같다. 첫째, AI VTuber 스트리밍의 지연 문제를 모델 처리 시간 문제가 아니라 지각된 반응성과 방송 연속성의 문제로 재정의한다. 둘째, 장문 응답 생성 중 발생하는 대화 공백을 완화하기 위한 데이터셋 근거 기반 사전 음성화 cover reaction을 연구 대상으로 제시한다. 셋째, 정서와 대화행위에 기반한 맥락 적합성이 체감 지연, 리액션 적절성, 대화 자연스러움, 사회적 존재감에 어떤 영향을 주는지 평가할 수 있는 관점을 제안한다.

2 관련 연구

Large Language Model (LLM)은 개방형 대화와 캐릭터 기반 응답 생성을 크게 확장했지만, 이를 아바타 기반 상호작용에 적

용할 때는 새로운 문제가 발생한다. LLM 기반 avatar chatbot 연구는 가상 아바타가 개방형 대화를 수행할 수 있음을 보였지만, 동시에 느린 생성 속도와 멀티모달 통합이 중요한 과제임을 지적하였다 [5]. 또한 generative agent 연구는 believable behavior를 위해 memory, reflection, planning과 같은 내부 상태 관리가 중요함을 보였으나, 이러한 구성요소는 실시간 방송형 시스템에서는 추가적인 처리 부담으로 이어질 수 있다 [8].

실시간 AI VTuber는 시청자의 채팅 입력을 해석하고, 캐릭터 페르소나에 맞는 언어 응답을 생성하며, 이를 음성과 아바타 표현으로 출력해야 한다. 텍스트 응답만 제공하는 환경에서는 수 초의 응답 지연이 비교적 허용될 수 있지만, 방송형 인터페이스에서는 캐릭터가 침묵하거나 멈춘 듯 보이는 순간 자체가 상호작용 품질 저하로 지각될 수 있다. 최근 LLM 기반 Embodied Conversational Agent (ECA) 연구에서도 low-latency multimodal response는 자연스러운 대화 흐름을 유지하기 위한 핵심 조건으로 다루어지고 있다 [6]. 따라서 AI VTuber의 지연 문제는 더 빠른 모델을 사용하는 성능 최적화 문제라기보다, 시청자가 대기 시간을 어떻게 지각하고 해석하는가의 문제로 확장된다.

응답 시간의 효과가 단순하지 않다는 점은 Human–chatbot interaction 연구에서도 확인된다. 관련 연구는 챗봇 응답 지연이 social presence와 usage intention에 서로 다른 방향으로 작용할 수 있으며, 사용자의 prior experience에 따라 같은 지연도 다르게 해석될 수 있음을 보였다 [9]. 이는 AI VTuber의 지연 완화가 물리적 대기 시간을 줄이는 것만으로 충분하지 않음을 시사한다. 즉, 시스템이 긴 응답을 아직 생성하고 있더라도, 그 사이에 캐릭터가 적절한 짧은 반응을 제공한다면 시청자는 이를 시스템 정지가 아니라 방송자가 듣고 반응을 준비하는 과정으로 해석할 가능성이 있다.

가상 에이전트와 디지털 휴먼 연구에서는 대화적 채움 발화(conversational filler), 제스처 기반 채움 행동(gestural filler), 다중양식 피드백(multimodal feedback)을 활용하여 응답 지연에 대한 부정적 인식을 완화하려는 시도가 이루어졌다. 지연된 가상 에이전트 응답 상황에서 대화적 채움 발화는 사용자가 시스템 지연을 인식하는 방식을 완화할 수 있는 전략으로 제안되었으며 [10], 디지털 휴먼과의 대화에서는 제스처 기반 채움 행동이 사용자의 체감 지연과 에이전트에 대한 인상 형성에 영향을 줄 수 있음이 보고되었다 [11]. 또한 LLM 기반 체화형 대화 에이전트 연구는 짧은 채움 발화, 비언어적 턴 전환 행동, 시각적 피드백과 같은 다중양식 피드백이 체감 지연과 몰입감에 영향을 줄 수 있음을 보였다 [7]. 자유 대화 상황의 LLM 기반 지능형 가상 에이전트 연구에서도 자연스러운 채움 반응이 긴 응답 지연의 부정적 효과를 완화할 수 있음이 제시되었다 [12]. 최근 연구 역시 대화적 · 제스처적 요소를 결합한 행동적 채움 반응이 진행 막대나 생각 말풍선과 같은 상징적 지시자보다 체감 응답 시간, 존재감, 인간다움, 자연스러움 측면에서 긍정적으로 작용할 수 있음을 보고하였다 [13].

그러나 기존 지연 완화 연구는 주로 일반 챗봇, 가상 에이전트, 디지털 휴먼, Virtual Reality (VR) 기반 ECA를 대상으로 하며, 공개 라이브 방송에서 AI VTuber가 장문의 응답을 준비하는 동안 어떤 종류의 짧은 반응을 먼저 제시해야 하는지는 충분히 다루지 않았다. VTuber 방송에서 짧은 반응은 단순한 로딩 표시나 “잠시만 기다려 주세요”와 같은 시스템 상태 알림으로만 기능하지 않는다. 시청자는 그 반응을 캐릭터의 성격, 방송 톤, 현재 채팅의 정서, 후속 발화와의 연결성, 그리고 아바타가 실제로 반응하고 있다는 사회적 단서로 해석한다. 또한 채움 발화가 항상 긍정적인 사용자 경험을 만드는 것은 아니다. Jeong et al.은 대화 에이전트가 “um”, “uh”와 같은 채움 발화를 사용할 때 사용자가 지각하는 지능성, 인간다움, 호감도가 어떻게 달라지는지를 분석하였다 [14]. 이는 채움 발화가 인간다운 말투를 흉내 내는 장치가 아니라, 대화 목적과 상황에 따라 긍정적 또는 부정적으로 해석될 수 있는 상호작용 요소임을 보여준다.

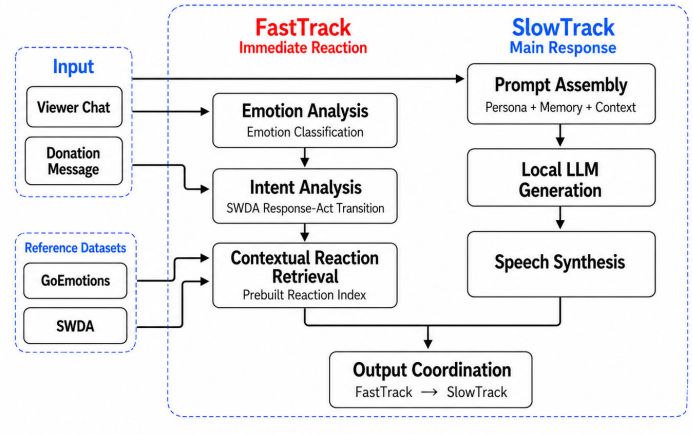

Figure 2: Dual-Track Response Generation Pipeline.

3 시스템 개요

Figure 2는 제안 시스템의 전체 처리 흐름을 보여준다. 제안 시스템은 AI VTuber가 장문 응답을 생성하는 동안 발생하는 침묵 구간을 완화하기 위한 dual-track latency-cover 구조이다. 핵심은 최종 응답이 준비될 때까지 시청자를 기다리게 하는 것이 아니라, 짧은 선행 반응을 먼저 제공하여 사용자가 캐릭터의 반응성을 지속적으로 지각하도록 하는 데 있다. 사용자 입력은 시청자 채팅과 도네이션 메시지로 구성된다. 입력이 발생하면 제안 시스템은 이를 하나의 장문 응답 생성 경로로만 처리하지 않고, 즉각적인 반응을 담당하는 FastTrack과 본문 응답을 담당하는 SlowTrack으로 나누어 처리한다.

본 구현은 Open-LLM-VTuber 런타임을 AI VTuber 상호작용 플랫폼으로 사용한다 [15]. FastTrack은 입력 직후 짧은 cover reaction을 선택한다. Emotion Analysis를 통해 발화의 정서적 뉘앙스를 분석하고, Intent Analysis를 통해 발화의 대화행위를 파악한다. 정서 분류에는 DistilBERT 기반 GoEmotions 분류기를 사용하고 [16], 대화행위 분류에는 SetFit 기반 접근을 사용한다 [17]. 또한 자연스러운 대응 행위 선택을 위해 Switchboard Dialogue

Act Corpus (SWDA)의 대화행위 정보를 참조하며, 이를 통해 사용자 발화의 의도를 그대로 반복하지 않고 적절한 response act로 전이한다 [18]. 이후 Contextual Reaction Retrieval 단계에서 정서 근거와 대화행위 근거를 활용하여 현재 입력에 적합한 짧은 cover reaction을 선택한다.

SlowTrack은 본문 응답을 생성하는 경로이다. SlowTrack은 캐릭터 페르소나, 최근 대화 맥락, 현재 입력을 바탕으로 프롬프트를 구성하고, 로컬 LLM을 통해 더 긴 본문 응답을 생성한 뒤 음성 합성 단계를 거쳐 출력으로 변환한다. 본 구현은 텍스트를 캐릭터 음성으로 변환하는 Style-Bert-VITS2 [19]와 Live2D 립싱크 출력 [20]을 조합한다. SlowTrack은 FastTrack보다 더 긴 처리 시간이 필요하지만, 정보 전달과 캐릭터 일관성을 담당하는 핵심 응답 경로이다.

FastTrack과 SlowTrack의 출력은 Output Coordination 단계에서 하나의 대화 턴으로 연결된다. 사용자는 먼저 짧은 FastTrack 반응을 경험하고, 이어서 SlowTrack의 본문 응답을 경험한다. 따라서 FastTrack은 최종 답변을 대체하는 경로가 아니라, 본문 응답이 준비되는 동안 방송 공백을 완화하고 캐릭터가 입력을 듣고 반응하고 있다는 신호를 제공하는 역할을 한다.

4 FastTrack

4.1 제안 목적 및 필요성

대규모 언어 모델과 음성 합성 기술을 결합한 가상 에이전트 연구가 활발히 진행됨에 따라, 이를 실시간 대화형 스트리밍에 적용하려는 시도가 지속되고 있다. 그러나 이러한 기법들에도 불구하고, AI VTuber 시스템에서 발생하는 물리적 지연 시간은 구조적으로 고착화될 수밖에 없다. 첫째, 실시간으로 수 시간 동안 지속되는 라이브 스트리밍의 특성상, 상용 대형 언어 모델 API를 지속적으로 호출하는 것은 경제적 지속 가능성(Economic Viability)을 담보하기 어렵다. 이로 인해 대다수 시스템은 로컬 자원을 활용한 온디바이스(On-device) 혹은 로컬 서버 배포 형식을 취하게 되며, 이는 로컬 하드웨어 성능에 종속된 컴퓨팅 지연을 발생시킨다.

둘째, 가상 캐릭터의 고유한 세계관과 엄격한 페르소나 규칙을 유지하기 위해서는 방대한 양의 시스템 프롬프트(System Prompt) 조율이 요구된다. 이에 더해 방송이 진행됨에 따라 시청자와의 대화 이력이 기억 장치(Memory System)에 지속적으로 축적되는데, 이러한 컨텍스트 윈도우(Context Window)의 확장은 LLM의 초기 토큰 생성 지연(Time-to-First-Token, TTFT)을 증가시키는 핵심 원인이 된다. 마지막으로, 상호작용의 질적 수준을 확보하기 위해 단답형 응답이 아닌 방송의 맥락에 부합하는 장문의 독백 및 서사 생성이 강제된다. LLM의 자동회귀적 디코딩(Autoregressive Decoding) 메커니즘 특성상 생성하는 문장이 길어질수록 전체 추론 시간은 누적되어 증가한다. 본 연구는 이러한 공백을 시스템 내부 연산의 한계 내에서만 최적화하는 대신, 초기 무응답 구간을 짧은 반응으로 완화하는 선행 경로로서 FastTrack을 제안한다.

4.2 FastTrack 구조

FastTrack은 사용자 입력 직후 대화의 정서적, 구조적 흐름을 실시간으로 분석하며 다음 모듈을 순차적으로 거친다. Emotion(DistilBERT/GoEmotions)은 사용자 발화에 내재된 정서적 뉘앙스를 분석하는 모듈이다. 실시간 방송 환경에서의 대기 시간을 최소화하고 안정적인 처리가 가능하도록 정서 라벨을 POSITIVE, NEGATIVE, SURPRISE, NEUTRAL의 네 가지 범주로 압축하여 판별한다. 다만 모델 파일이 아직 로드되지 않았거나 분류 호출이 실패하는 경우에는 시스템이 중단되지 않도록, 감정 키워드와 문장부호를 이용한 단순 규칙으로 임시 정서 라벨을 반환한다.

Intent(SWDA/SetFit)는 사용자 발화의 의도와 대화행위를 파악하는 분류 모듈이다. 입력된 문장을 QUESTION, INFORM, ACKNOWLEDGE, DIRECTIVE, EXPRESSIVE, REJECT의 여섯 가지 핵심 라벨로 구조화한다. Response-Act Transition은 분석된 사용자 의도에 대해 시스템이 취해야 할 대답 행동을 결정하는 전이 모듈이다. 이 모듈은 사용자의 입력 대화행위(Incoming intent)와 시스템의 출력 반응(Output response act)을 분리하고, 질문 뒤에는 정보 제공이나 확인 반응이, 지시 뒤에는 수락이나 거절 반응이 이어지는 것처럼 인접 발화 쌍(Adjacency Pairs)의 대응 가능성을 전이 행렬로 표현한다. 런타임에서는 분류된 입력 행위에 해당하는 행을 조회하여 가장 적절한 출력 행위를 선택함으로써 자연스러운 상호작용 흐름을 유도한다.

4.3 데이터셋

본 연구에서 제안하는 FastTrack 계층의 핵심 분류 및 검색 기반은 자연어 처리 분야에서 검증된 표준 영어 데이터셋에 의존한다. 감정 분석의 기준이 되는 GoEmotions는 Reddit 커뮤니티의 데이터로부터 추출된 영어 기반 데이터셋이다 [16]. 또한, 의도 분류 및 전이의 기준이 되는 SWDA는 인간 간의 자발적인 대화를 기록한 음성 전사 코퍼스인 Switchboard를 바탕으로 구축된 대표적인 영어 대화 데이터셋이다 [18]. 이처럼 가상 에이전트의 즉각적인 언어 · 정서 반응을 제어하기 위한 핵심 모듈들이 영어 텍스트 자산에 근거하고 있으므로, 본 연구의 Emotion, Intent, Contextual Reaction Retrieval 파이프라인은 영어권의 컨텍스트 하에서 일관되게 작동하도록 설계되었다.

검색 대상 리액션 자산은 원자료, 검색 근거, 페르소나 표면 발화의 세 단계로 구축하였다. GoEmotions 원 CSV 211,225행을 네 범주로 통합한 전처리 데이터는 43,410문장(POSITIVE 12,591, NEGATIVE 8,027, SURPRISE 3,693, NEUTRAL 19,099)으로 구성된다. SWDA의 경우 전처리 후 145,004개 발화(QUESTION 6,823, INFORM 92,952, ACKNOWLEDGE 41,280, DIRECTIVE 705, EXPRESSIVE 2,761, REJECT 483)로 구성된다.

두 데이터에 결정론적 규칙 필터와 LLM 품질 검사를 순차 적용하였다. 질문형과 QUESTION 대화행위, URL, 숫자 · 날짜, 고유명사 · 브랜드 · 플랫폼명, 비속어와 폭력적 · 성적 표현, 이모지 · 감탄 소음, 인사 · 종결 표현, 미완성 · 문맥 의존 문장, 14단어 또는 96자 초과 문장 및 중복을 제거하였다. 최종 검색 근거 풀은 GoEmotions 385개(POSITIVE 141, NEGATIVE 133, SURPRISE 23, NEUTRAL 88)와 SWDA 244개(INFORM 199, ACKNOWLEDGE 3, DIRECTIVE 18, EXPRESSIVE 14, REJECT 10), 총 629개이다. 이는 전체 출력 반응 수가 아닌 검색 · 근거 추적용 원천 문장 수이다. 두 풀은 출처별로 분리 저장하며, Grounded 조건은 정서와 대화행위 근거의 교차 공간(Cross-pair space)을 탐색한다.

이 629개 근거 문장은 런타임에 그대로 출력되는 문장이 아니라, 복수의 페르소나 구어체 표면 발화로 확장된다. 확장된 표면 발화는 길이, 질문형, 반복 및 페르소나 부적합 여부를 다시 검사한 뒤 Style-Bert-VITS2로 오디오 파일(.wav)을 사전 생성한다. 여기서 사전 음성화 매니페스트는 음성 합성 모델 자체가 아니라, 각 표면 발화의 ID와 미리 생성된 WAV 파일을 연결하는 JSON 기반 색인 파일이다. 매니페스트의 각 항목에는 원천 근거 문장, 페르소나 표면 발화, 데이터 출처, 정서 또는 대화행위 라벨, 오디오 파일 경로가 기록된다. 본 실험의 매니페스트는 총 1,816개 항목(GoEmotions 계열 917개, SWDA 계열 899개)을 포함하므로, 629개 근거 문장과 1,816개 실제 출력 자산의 생성 관계를 역추적할 수 있다.

4.4 Contextual Reaction Retrieval

Contextual Reaction Retrieval은 앞선 단계에서 도출된 정서적 뉘앙스와 대응 행동 정보를 결합하여, 시스템이 출력할 최적의 선행 반응 대사(Text Asset)를 동적으로 검색하는 모듈이다. 본 모듈은 Emotion Analysis 모듈에서 추정된 4대 정서 라벨과 Intent Analysis 모듈 및 Response-Act Transition을 통해 계산된 출력 대화행위를 입력값으로 받아 일치하는 후보군을 조회한다. 이때 선행 반응이 다시 질문형으로 끝나 후속 SlowTrack 본문 답변과 문맥상 충돌하는 현상을 방지하기 위해, QUESTION 형태는 검색 대상 행위에서 제외된다.

검색 절차는 네 단계로 구성된다. 첫째, 입력 발화를 Emotion Analysis 모듈과 Intent Analysis 모듈에 동시에 전달하여 정서 라벨과 incoming dialogue act를 얻는다. 둘째, SWDA 기반 전이 행렬을 사용해 incoming dialogue act를 시스템이 출력할 response act로 변환한다. 셋째, 실험 조건에 따라 정서 라벨이 가리키는 GoEmotions 후보 공간, response act가 가리키는 SWDA 후보 공간 또는 두 근거를 함께 사용하는 후보 공간을 조회한다. 넷째, 해당 조건의 후보 중 하나를 선택하고, 선택된 항목의 ID를 사전 음성화 매니페스트와 매칭하여 재생할 오디오 경로를 결정한다. 이 방식은 사용자의 입력 의도를 그대로 반복하지 않고, 조건별 정서 및 대화행위 근거에 대응하는 선행 반응을 즉시 출력하도록 설계되었다.

4.5 Prebuilt WAV Manifest Match 및 Prebuilt Audio Match

앞 절에서 정의한 사전 음성화 매니페스트는 런타임에서 FastTrack 텍스트와 실제 오디오 파일을 연결하는 조회 테이블로 사용된다. 즉, Contextual Reaction Retrieval은 현재 조건에 맞는 텍스트 ID를 결정하고, Prebuilt WAV Manifest Match와 Prebuilt Audio Match는 해당 ID를 재생 가능한 음성 자산으로 변환한다.

런타임에 검색된 FastTrack 텍스트가 확정되면 시스템은 두 가지 매칭 모듈을 통해 음성을 즉시 출력한다. Prebuilt WAV Manifest Match는 선택된 텍스트 ID를 기반으로 사전에 구축된 오디오 매니페스트를 조회하고, 해당 텍스트에 매핑된 구체적인 파일 경로와 데이터 출처 메타데이터를 찾아낸다. Prebuilt Audio Match는 매칭된 경로 정보를 바탕으로 해당 사전 생성 오디오 파일을 시스템 오디오 렌더러에 즉각 연결하여 재생을 수행한다. 이러한 파이프라인 구조를 통해 FastTrack은 입력 문장의 길이와 관계없이 감정 · 의도 분류 및 인덱스 참조에 필요한 최소한의 연산만 수행하므로, 런타임 TTS 지연을 배제한 채 안정적인 처리 속도로 초기 무응답 구간을 메울 수 있다.

4.6 FastTrack 처리 지연 측정

Table 1: FastTrack input-length-wise mean module latency (ms).

Input length bin	n	Emotion	Intent	Retrieval	Total
10–100 chars	30	10.8	9.2	16.7	36.7
100–200 chars	30	10.8	9.5	16.1	36.4
200–500 chars	30	14.2	10.6	19.2	44.0
500–1000 chars	30	22.0	15.6	16.3	53.9

Table 1은 입력 텍스트 길이에 따른 FastTrack 경로의 모듈별 평균 처리 시간을 나타낸다. 각 입력 길이 구간에 대해 30회 반복 측정을 수행하였으며, 표의 값은 각 모듈의 평균 처리 지연시간(ms)을 의미한다. 측정 결과, FastTrack의 전체 처리 시간은 모든 입력 길이 구간에서 100 ms 이하로 유지되었다. 10–100자 구간에서는 36.7 ms, 100–200자 구간에서는 36.4 ms, 200–500자 구간에서는 44.0 ms, 500–1000자 구간에서는 53.9 ms로 측정되었다. 이는 입력 길이가 증가하더라도 FastTrack이 실시간 상호작용에 적합한 짧은 처리 시간을 유지함을 보여준다.

5 SlowTrack 및 스케줄링

SlowTrack은 캐릭터가 실제로 말할 본문 응답을 생성하는 경로이다. 로컬 LLM은 시스템 프롬프트, 현재 채팅 batch, 최근 대화 memory를 바탕으로 응답을 생성한다. memory는 장기 semantic memory가 아니라 JSON 기반의 경량 프로필, 최근 턴, 감정 이벤트 저장소로 설계한다. 이는 latency를 크게 늘리지 않는 범위에서 최소한의 맥락 지속성을 제공하기 위한 대안이다.

Table 2: SlowTrack output-length-wise mean module latency (ms).

Output length bin	n	LLM gen.	TTS synth.	Total
10–100 chars	30	355.0	350.7	705.7
100–200 chars	30	608.0	511.3	1,119.3
200–500 chars	30	1,232.9	857.0	2,089.9
500–1000 chars	30	2,501.8	1,618.8	4,120.6

Table 2는 SlowTrack 경로에서 출력 길이 구간별로 발생한 평균 모듈 지연 시간을 나타낸다. SlowTrack은 로컬 LLM(qwen2.5:7b)을 통해 본문 응답을 생성한 뒤 Style-Bert-VITS2 기반 실시간 TTS 합성을 수행하므로, 출력 길이가 증가할수록 LLM 생성 시간과 TTS 합성 시간이 함께 증가하는 경향을 보인다. 측정 결과, 10–100자 구간의 총 평균 지연은 705.7 ms였으나 500–1000자 구간에서는 4,120.6 ms까지 증가하였다. 이는 SlowTrack이 장문 응답과 캐릭터 일관성을 제공하는 핵심 응답 경로인 동시에, 실시간 방송 환경에서는 시청자가 체감할 수 있는 침묵 구간을 발생시킬 수 있음을 보여준다.

SlowTrack 프롬프트는 단순 챗봇 설정이 아니라 방송형 VTuber의 실제 생성 부담을 반영하도록 확장된다. 프롬프트에는 VTuber 페르소나, 대학원 고민 상담과 같은 방송 콘텐츠, 최근 도네이션 및 채팅 요약, 시청자별 반복 주제, 진행 중인 방송 세그먼트, 금지된 필러 및 캐치프레이즈 규칙이 함께 포함된다. 따라서 SlowTrack 지연은 임의로 삽입한 대기 시간이 아니라, 페르소나 일관성, 대상 시청자 인식, 방송 흐름을 유지하기 위한 문맥 부하의 일부로 정의된다.

스케줄링은 Serial을 주 실험 조건으로 사용한다. 현재 발화가 끝난 뒤 다음 입력 처리와 응답 생성을 시작하므로, 사전 음성화 FastTrack이 실제 침묵 구간을 얼마나 줄이는지 직접 관찰할 수 있다. 도네이션 이벤트는 별도의 순차 전달 절차를 갖는다. 브라우저는 먼저 도네이션 overlay와 readout TTS를 재생하고, readout 재생이 완료된 이후에만 VTuber 입력을 전달한다. 여기서 순차 전달 절차는 후원 메시지 낭독과 캐릭터 응답이 겹치지 않도록 두 이벤트 사이에 재생 완료 조건을 두는 것을 의미한다. 이는 실제 방송에서 후원 음성을 들은 뒤 캐릭터가 반응하는 절차를 모사하기 위한 것이다.

6 실험 방법

6.1 참가자 및 조건 매핑

총 32명의 유효 설문 응답을 수집하였다. 본 실험은 동일한 참가자가 모든 실험 조건을 평가하는 within-subjects 설계로 진행되었다. 조건명에 따른 평가 편향을 줄이기 위해 영상에는 조건명을 노출하지 않았으며, 모든 참가자에게 Intent Only, Grounded, SlowTrack Only, Emotion Only, Neutral Random 순서로 제시하였다. 모든 문항은 7점 Likert 척도로 측정하였다. 부정 문항인 침묵/공백 답답함 문항과 무작위 반응 지각 문항은 점수가 높을수록 긍정적인 평가가 되도록 역코딩하였다.

6.2 평가지표

본 연구의 평가지표는 AI VTuber의 latency-cover 반응이 단순히 물리적 응답 시간을 줄이는 것이 아니라, 시청자가 대기 시간을 어떻게 지각하고 해석하는지를 측정하기 위해 구성하였다. Perceived Latency는 LLM 기반 intelligent virtual agent의 free-form conversation에서 conversational filler가 perceived response time을 완화할 수 있음을 보인 Maslych et al.의 연구를 참조하여, 사용자 채팅 이후 응답 대기 시간이 얼마나 짧고 자연스럽게 느껴졌는지를 측정한다 [12]. Reaction Appropriateness는 filler의 형태와 맥락이 사용자 인상과 체감 지연에 영향을 줄 수 있다는 선행 연구를 바탕으로, FastTrack 반응이 단순한 시간 끌기가 아니라 사용자 발화의 정서와 대화 맥락에 맞는 의미 있는 반응으로 지각되었는지를 평가한다 [11, 14]. Conversational Naturalness는 짧은 언어적 · 비언어적 피드백이 체화형 대화 에이전트의 상호작용 경험과 몰입에 영향을 줄 수 있다는 연구를 참고하여, FastTrack과 SlowTrack을 포함한 전체 대화 흐름이 실제 방송 상황처럼 자연스럽게 이어졌는지를 측정한다 [7, 12].

평가 구성개념과 각 구성개념의 측정 초점은 Table 3와 같다.

Table 3: Subjective constructs and measurement focus.

Construct	Measurement focus	Refs.
Perceived Latency	사용자 채팅 이후 기다리는 시간이 짧고 자연스럽게 느껴졌는가	[12]
Reaction Appropriateness	짧은 리액션이 사용자 발화의 정서와 대화 맥락에 맞는가	[11,14]
Conversational Naturalness	FastTrack과 SlowTrack을 포함한 전체 대화 흐름이 실제 방송처럼 자연스러운가	[7]
Social Presence	AI VTuber가 실제로 듣고 반응하는 사회적 존재처럼 느껴지는가	[21]
Character Appeal	캐릭터의 매력과 리액션을 통한 캐릭터성 강화가 지각되는가	[22]
Engagement	대화 상황에 몰입하고 AI VTuber의 반응을 계속 보고 싶은가	[1,3]

또한 Social Presence는 온라인 상호작용에서 타자가 실제로 존재하는 것처럼 지각되는 정도를 개념화한 Kreijns et al.의 연구를 바탕으로, AI VTuber가 단순한 음성 출력 장치가 아니라 시청자의 말을 듣고 상호작용하는 사회적 존재로 받아들여지는지를 평가한다 [21]. Character Appeal은 Godspeed Questionnaire의 Likeability와 Animacy를 참고해 캐릭터 매력과 리액션에 의한 캐릭터성 강화를 측정한다 [22]. Engagement는 VTuber 시청 경험의 관여와 관계 형성 논의를 참고해 영상 몰입과 후속 반응의 지속 시청 의도를 측정한다 [1, 3]. 따라서 이 세 척도는 선행 척도의 단순 번안이 아니라 본 실험 목적에 맞게 조작화한 척도이다.

Table 4: Subjective evaluation means and standard deviations by condition for all participants.

Mean (SD)	Grounded	Emotion Only	Intent Only	Neutral Random	SlowTrack Only
Overall	4.95 (1.11)	4.91 (1.24)	4.81 (0.95)	4.69 (1.30)	4.48 (1.28)
Perceived Latency	5.47 (1.00)	5.24 (1.33)	5.26 (1.05)	5.12 (1.03)	4.71 (1.42)
Reaction Appropriateness	4.83 (1.10)	4.79 (1.25)	4.93 (0.89)	4.73 (1.33)	4.74 (1.19)
Conversational Naturalness	4.75 (1.49)	4.84 (1.66)	4.58 (1.29)	4.30 (1.47)	4.25 (1.66)
Social Presence	5.30 (1.24)	5.27 (1.31)	5.31 (1.38)	4.86 (1.53)	4.78 (1.55)
Character Appeal	4.44 (1.57)	4.67 (1.73)	4.05 (1.54)	4.56 (1.83)	4.09 (1.79)
Engagement	4.72 (1.53)	4.53 (1.67)	4.45 (1.42)	4.33 (1.78)	4.05 (1.75)

7 실험 결과

7.1 설문 척도 신뢰도

32명의 5개 조건별 응답 160건을 통합하여 14개 문항을 하나의 종합 척도로 보았을 때 Cronbach's alpha는 .93으로 높게 나타났다. 하위 척도에서는 Social Presence(.916), Character Appeal(.846), Engagement(.799), Conversational Naturalness(.790)가 비교적 양호한 신뢰도를 보였다. Perceived Latency는 .693으로 수용 가능한 수준에 근접하였다. 반면 Reaction Appropriateness는 .557로 상대적으로 낮게 나타났으므로, 해당 지표는 평균 점수만으로 강하게 해석하기보다 최종 선택 문항과 함께 보조적으로 해석하였다.

7.2 조건별 기술통계

Table 4는 전체 응답자 32명을 대상으로 각 조건별 주관 평가의 평균과 표준편차를 정리한 것이다.

Table 4의 기술통계 경향성을 살펴보면, FastTrack을 구성하는 두 핵심 요인인 의도(Intent)와 감정(Emotion)이 사용자 경험의 서로 다른 차원에 기여하고 있음을 보여주는 기능적 분화 경향이 관찰된다. Intent Only 조건은 반응 적절성(Reaction Appropriateness, M=4.93) 지표에서 전체 조건 중 가장 높은 평균을 기록하였다. 반면, Emotion Only 조건은 캐릭터 매력도(Character Appeal, M=4.67)뿐만 아니라 대화 자연스러움(Conversational Naturalness, M=4.84)에서도 대조군을 포함한 모든 조건 중 가장 높은 점수를 기록하였다. 이는 의도 정보가 대화의 구조적 정합성을 견인하는 반면, 감정 정보는 에이전트의 정서적 자연스러움과 사회적 호감도를 지탱하는 기제로 작동할 가능성을 시사한다.

7.3 조건 간 차이에 대한 Omnibus 검정

5개 조건 전체의 차이를 확인하기 위해 Friedman test를 수행하였다. 주요 결과는 Table 5와 같다.

Friedman test 결과, 대부분의 지표에서는 조건 간 차이가 통계적으로 유의하지 않았으나, Character Appeal에서는 유의한 조건 차이가 관찰되었다($\chi^2(4) = 13.637$, $p = 0.009$, Kendall's W = .107). Perceived Latency는 $p = 0.062$, Conversational Naturalness는 $p = 0.083$으로 유의수준에 근접한 경향을 보였으나 0.05 수준에서는 유의하지 않았다. 따라서 본 결과는 특정 조건의 전면적 우수성을 입증하기보다는 FastTrack 기반 latency-cover 구조가 일부 평가 지표에서 긍정적인 신호를 보였다는 탐색적 결과로 해석하는 것이 적절하다.

Table 5: Friedman test results by condition for all participants.

Measure	χ^2	df	p	Kendall's W
Overall	7.373	4	.117	.058
Perceived Latency	8.979	4	.062	.070
Conversational Naturalness	8.233	4	.083	.064
Reaction Appropriateness	0.808	4	.937	.006
Social Presence	5.488	4	.241	.043
Character Appeal	13.637	4	.009	.107
Engagement	4.647	4	.326	.036

7.4 SlowTrack Only 대비 계획 비교

계획 비교(Planned Comparison)를 통해 대조군인 SlowTrack Only 대비 각 FastTrack 조건의 효과를 paired t-test로 검증한 결과, 감정 기반 선행 반응이 사용자 지각에 미치는 다면적인 효과가 관찰되었다. Emotion Only 조건은 캐릭터 매력도(Character Appeal)에서 SlowTrack Only보다 유의하게 높은 평가를 받았다(mean diff = 0.58, $t(31) = 2.71$, $p = 0.011$, Holm 보정 $p = 0.043$, $d_z = .48$). 이는 대화의 구조적 맥락이 다소 결여되더라도, 사용자의 감정 상태에 즉각적으로 동조하는 정서적 선행 반응이 가상 캐릭터의 생동감과 호감도 형성에 긍정적으로 작용할 수 있음을 보여준다.

Conversational Naturalness 지표의 경우, Emotion Only 조건은 기술통계상 최고점을 기록하였으며 SlowTrack Only 조건과의 paired t-test에서도 보정 전 p값 기준으로는 유의한 차이를 나타냈다(mean diff = 0.594, $t(31) = 2.289$, $p = .029$, $d_z = 0.405$). 다만 동일 지표 내 4개 비교에 대한 Holm 보정을 적용할 경우 유의수준에 도달하지 못하였다(Holm $p = .116$). 따라서 전체 표본 기준에서는 감정 기반 FastTrack이 대화의 자연스러움을 향상시키는 긍정적인 경향성을 보였다고 해석하는 것이 적절하다.

7.5 영어 친숙 하위집단 탐색 분석

본 연구의 FastTrack 파이프라인을 구성하는 핵심 자산인 감정 분류 모듈(GoEmotions)과 대화행위 전이 모듈(SWDA)은 영어권 언어 패턴과 정서적 리액션 구조를 기반으로 구축된 표준 영어 데이터셋이다. 이에 따라 실험에 사용된 5가지 조건의 자극 영상 역시 시스템의 언어적·문화적 컨텍스트를 보존하기 위해 영어 상호작용 환경으로 제작되었다.

이러한 실험 자극의 언어적 특성이 사용자 지각에 미치는 영향을 보조적으로 살펴보기 위해, 영어 기반 가상 캐릭터(VTuber) 콘텐츠 시청 빈도가 높은(7점 척도 중 4점 이상) 참가자 집단만을 분리한 사후적(post-hoc) 탐색 하위 표본 분석을 실시하였다(n=19). 분석 결과, 영어 콘텐츠 시청 경험이 풍부한 하위 집단에서는 Emotion Only 조건이 SlowTrack Only 조건보다 대화 자연스러움(Conversational Naturalness)에서 Holm 보정을 적용한 후에도 높은 점수를 기록하였다(mean diff = 1.079, $t(18) = 3.021$, $p = .007$, Holm $p = .029$, $d_z = 0.693$). 다만 이 분석은 사전에 확정된 주 검정이 아니라 언어 친숙도 가능성을 살피기 위한 사후 탐색 분석이므로, 결과는 후속 연구에서 검증해야 할 가설 생성적 근거로 해석한다.

7.6 최종 선택 문항 결과

복수 선택이 가능한 최종 비교 문항에서 가장 대화 흐름이 자연스러웠던 영상으로는 Intent Only 조건이 22명에게 선택되어 가장 높았고, Grounded 조건이 20명으로 뒤를 이었다. 리액션이 맥락에 가장 잘 어울렸던 영상 역시 Intent Only 조건이 23명, Grounded 조건이 21명에게 선택되었다. 반면 가장 어색하거나 불편했던 영상으로는 SlowTrack Only 조건이 14명에게 선택되어 가장 높은 빈도를 보였다.

7.7 결과 요약

종합하면, 본 실험 결과는 FastTrack 기반 latency-cover 접근법의 가능성과 각 구성 모듈의 상호보완적 역할을 탐색적으로 보여준다. 정량적 지연 시간 분석이 FastTrack의 신속한 초기 반응 경로를 보여준다면, 주관적 설문 결과는 두 모듈이 서로 다른 방식으로 사용자 경험에 기여할 가능성을 시사한다. Intent 기반 모듈은 대화의 구조적 흐름과 맥락 적합성을 통제하여 최종 선택 문항에서의 높은 선호도와 관련될 수 있으며, Emotion 기반 모듈은 캐릭터의 매력도와 대화의 정서적 자연스러움을 지탱하는 요인으로 작용할 수 있다. 다만 대부분의 omnibus 검정이 유의하지 않았고, 영어 친숙 하위집단 분석은 사후 탐색 분석이므로, 본 결과는 확정적 우위의 증거라기보다 후속 실시간 실험을 위한 설계 근거로 해석해야 한다.

8 결론

본 연구는 AI VTuber 스트리밍에서 발생하는 응답 지연 문제를 단순한 시스템 처리 시간의 문제가 아니라, 시청자가 방송 중 대기 시간을 어떻게 지각하고 해석하는가의 문제로 재정의하였다. 이를 위해 즉각적인 선행 반응을 담당하는 FastTrack과 장문 응답을 담당하는 SlowTrack을 분리한 dual-track latency-cover 구조를 제안하였다. FastTrack은 사용자 발화의 정서와 대화행위를 분석하고, 사전 음성화된 짧은 cover reaction을 즉시 선택함으로써 SlowTrack 응답이 준비되는 동안 발생하는 침묵 구간을 완화하도록 설계되었다.

시스템 성능 측정에서 FastTrack은 입력 길이가 증가하더라도 모든 구간에서 100 ms 이하의 짧은 평균 처리 시간을 유지하였으며, 이는 실시간 방송 환경에서 초기 반응 경로로 활용될 수 있음을 보여준다. 반면 SlowTrack은 로컬 LLM 응답 생성과 실시간 TTS 합성을 포함하기 때문에 출력 길이가 증가할수록 지연 시간이 크게 증가하였다. 사용자 설문 실험에서도 FastTrack 기반 조건은 SlowTrack Only 조건에 비해 여러 지표에서 긍정적인 평균 경향을 보였으며, 특히 Character Appeal에서는 유의한 조건 차이가 관찰되었다. 그러나 대부분의 지표에서 omnibus 검정은 유의하지 않았으므로, 본 연구는 특정 FastTrack 조건의 전면적 우수성을 단정하지 않는다.

본 연구의 한계도 명확하다. 참가자들이 평가한 것은 실제 양방향 라이브 방송이 아니라 사전 제작된 영상 자극이었으며, 따라서 실제 스트리밍 상황에서 발생하는 입력 변동성, 네트워크 지연, 시청자 동시성, 예측 불가능한 방송 흐름은 충분히 반영되지 않았다. 또한 모든 참가자에게 조건을 동일한 고정 순서로 제시했으므로 순서 효과와 피로 효과를 완전히 배제할 수 없다. FastTrack 반응 풀은 필터링과 사전 음성화를 통해 구성되었지만, 반응 다양성, 장기 반복 노출 효과, 동일 반응의 반복 노출 가능성은 더 체계적으로 검증되어야 한다. 본 실험은 SlowTrack Only와 FastTrack mapping 조건 간 비교에 초점을 두었으므로, conversational filler, gestural filler, multimodal feedback, 로딩 표시와 같은 기존 지연 완화 baseline과의 직접 비교는 포함하지 못했다. 따라서 본 결과는 제안 방법이 모든 지연 완화 방식보다 우수하다는 결론이 아니라, AI VTuber 맥락에서 데이터셋 근거 기반 선행 반응을 평가할 수 있는 초기 실험 근거로 해석해야 한다.

향후 연구에서는 더 많은 참가자와 다양한 방송 시나리오를 대상으로 실험을 확장하고, 실제 실시간 상호작용 환경에서 FastTrack 반응의 효과를 확인할 필요가 있다. 정서 기반 반응, 의도 기반 반응, Grounded 반응이 대화 유형에 따라 어떻게 다르게 지각되는지 분석함으로써, AI VTuber의 반응 선택 전략을 보다 정교하게 설계할 수 있을 것이다. 최종적으로 본 연구는 AI VTuber의 지연 문제를 물리적 속도 개선만으로 해결하려는 접근과 달리, 시청자의 체감 반응성과 사회적 존재감을 유지하는 상호작용 설계 문제로 확장했다는 점에서 의의를 가진다.

References

[1] Z. Lu, C. Shen, J. Li, H. Shen, and D. Wigdor, "More kawaii than a real-person live streamer: Understanding how the otaku community engages with and perceives virtual youtubers," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–14, doi: 10.1145/3411764.3445660.

[2] Q. Wan and Z. Lu, "Investigating vtubing as a reconstruction of streamer self-presentation: Identity, performance, and gender," *Proceedings of the ACM on Human-Computer Interaction*, vol. 8, no. CSCW1, pp. 1–22, 2024, doi: 10.1145/3637357.

[3] M. Yin, C. Shen, and R. Xiao, "Entertainers between real and virtual - investigating viewer interaction, engagement, and relationships with avatarized virtual livestreamers," in *Proceedings of the 2025 ACM International Conference on Interactive Media Experiences*, 2025, pp. 243–257, doi: 10.1145/3706370.3727866.

[4] Y. Hu and G. Freeman, "Beyond a conventional chatbot: How ai streamers transcend live streaming experiences from viewers' perspectives," in *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 2026, pp. 1–18, doi: 10.1145/3772318.3791077.

[5] T. Yamazaki, T. Mizumoto, K. Yoshikawa, M. Ohagi, T. Kawamoto, and T. Sato, "An open-domain avatar chatbot by exploiting a large language model," in *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2023, pp. 428–432, doi: 10.18653/v1/2023.sigdial-1.40.

[6] K. W. Kühlem, J. Ehret, T. W. Kuhlen, and A. Bönsch, "A latency-optimized llm-based multimodal dialogue system for embodied conversational agents in vr," in *Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*, 2025, pp. 1–3, doi: 10.1145/3717511.3749287.

[7] M. Elfleet and M. Chollet, "Investigating the impact of multimodal feedback on user-perceived latency and immersion with llm-powered embodied conversational agents in virtual reality," in *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents*, 2024, pp. 1–9, doi: 10.1145/3652988.3673965.

[8] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023, pp. 1–22, doi: 10.1145/3586183.3606763.

[9] U. Gnewuch, S. Morana, M. T. P. Adam, and A. Maedche, "Opposing effects of response time in human–chatbot interaction," *Business & Information Systems Engineering*, vol. 64, no. 6, pp. 773–791, 2022, doi: 10.1007/s12599-022-00755-x.

[10] H.-A. Boukaram, M. Ziadee, and M. F. Sakr, "Mitigating the effects of delayed virtual agent response time using conversational fillers," in *Proceedings of the 9th International Conference on Human-Agent Interaction*, 2021, pp. 130–138, doi: 10.1145/3472307.3484181.

[11] J. Kum and M. Lee, "Can gestural filler reduce user-perceived latency in conversation with digital humans?" *Applied Sciences*, vol. 12, no. 21, 2022, doi: 10.3390/app122110972.

[12] M. Maslych, M. Katebi, C. Lee, Y. Hmaiti, A. Ghasemaghaei, C. Pumarada, J. Palmer, E. S. Martinez, M. Emporio, W. Snipes, R. P. McMahan, and J. J. LaViola Jr., "Mitigating response delays in free-form conversations with llm-powered intelligent virtual agents," in *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, 2025, pp. 1–15, doi: 10.1145/3719160.3736636.

[13] D. M. Gonzales, S. Kalamkar, S. Jörg, and J. Grubert, "Behavioral and symbolic fillers as delay mitigation for embodied conversational agents in virtual reality," *IEEE Transactions on Visualization and Computer Graphics*, vol. 31, no. 11, pp. 9783–9791, 2025, doi: 10.1109/TVCG.2025.3616865.

[14] Y. Jeong, J. Lee, and Y. Kang, "Exploring effects of conversational fillers on user perception of conversational agents," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–6, doi: 10.1145/3290607.3312913.

[15] Open-LLM-VTuber Project, "Open-llm-vtuber," GitHub repository, 2026, accessed: 2026-06-11. [Online]. Available: https://github.com/Open-LLM-VTuber/Open-LLM-VTuber

[16] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "Goemotions: A dataset of fine-grained emotions," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4040–4054, doi: 10.18653/v1/2020.acl-main.372.

[17] L. Tunstall, N. Reimers, U. E. S. Jo, L. Bates, D. Korat, M. Wasserblat, and O. Pereg, "Efficient few-shot learning without prompts," arXiv:2209.11055, 2022, doi: 10.48550/arXiv.2209.11055.

[18] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, vol. 26, no. 3, pp. 339–373, 2000, doi: 10.1162/089120100561737.

[19] litagin02, "Style-bert-vits2," GitHub repository, 2026, accessed: 2026-06-11. [Online]. Available: https://github.com/litagin02/Style-Bert-VITS2

[20] Live2D Inc., *Lip-sync*, Live2D Cubism SDK Manual, 2026, accessed: 2026-06-11. [Online]. Available: https://docs.live2d.com/en/cubism-sdk-manual/lipsync/

[21] K. Kreijns, K. Xu, and J. Weidlich, "Social presence: Conceptualization and measurement," *Educational Psychology Review*, vol. 34, no. 1, pp. 139–170, 2022, doi: 10.1007/s10648-021-09623-8.

[22] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *International Journal of Social Robotics*, vol. 1, no. 1, pp. 71–81, 2009, doi: 10.1007/s12369-008-0001-3.

〈저자소개〉

이 영 성

- 2024 한동대학교 ICT창업학부 공학사
- 2025 - 현재 한동대학교 AI융합학부 휴먼&테크AI융합학과 석사 과정
- 관심 분야: 컴퓨터그래픽스, VR/AR, HCI, 게임공학
- https://orcid.org/0009-0006-2310-8008

최 재 효

- 2020 - 현재 한동대학교 AI융합학부 재학 중
- 관심 분야: 컴퓨터그래픽스, 인공지능 딥러닝, HCI, 디지털 트윈
- https://orcid.org/0009-0005-5397-3265

김 주 원

- 2023 한동대학교 ICT창업학부 공학사
- 2025 (주)이콜트리 인턴 2개월
- 2026 - 현재 한동대학교 AI융합학부 휴먼&테크AI융합학과 석사 과정
- 관심 분야: 컴퓨터그래픽스, 인공지능 딥러닝, 유체 시뮬레이션, 물리 기반 시뮬레이션, 디지털 트윈
- https://orcid.org/0009-0006-3606-0584

김 세 하

- 2025 - 현재 한동대학교 AI융합학부 ICT융합전공 재학 중
- 관심 분야 : 컴퓨터그래픽스, 디지털 콘텐츠, 비주얼 미디어, 캐릭터 애니메이션
- https://orcid.org/0009-0007-3680-339X

한 다 성

- 2006년 광운대학교 컴퓨터공학부 컴퓨터소프트웨어전공 학사 2008년 한국과학기술원 전자전산학과 전산학전공 석사 2014년 한국과학기술원 전산학과 박사
- 2014년 - 2016년 카이스트 문화기술 연구소 박사후 연구원 2016년 - 2022년 한동대학교 ICT창업학부 조교수
- 2023년 - 2025년 한동대학교 ICT창업학부 부교수
- 2026년 - 현재 한동대학교 AI융합학부 부교수
- 관심분야: 캐릭터 애니메이션, 물리 기반 시뮬레이션, 동작 제어, VR/AR
- https://orcid.org/0000-0003-1455-5114