Congruency-Dependent Gaze Allocation to Avatar Facial Expressions during Shared Video Viewing in Immersive VR

Joshua Smallberg
Computer Science
University of Bonn
Bonn, Germany
s6josmal@uni-bonn.de

Sina Kitz
Computer Science
University of Bonn
Bonn, Germany
kitz@cs.uni-bonn.de

Jan Spindler
Computer Science
University of Bonn
Bonn, Germany
jspindle@uni-bonn.de

Bardia Ardakanian
Computer Science
University of Bonn
Bonn, Germany
s55barda@uni-bonn.de

Leif Van Holland
Computer Science
University of Bonn
Bonn, Germany
holland@cs.uni-bonn.de

Nils Wandel
Computer Science
University of Bonn
Bonn, Germany
wandeln@cs.uni-bonn.de

Reinhard Klein
Computer Science
University of Bonn
Bonn, Germany
rk@cs.uni-bonn.de

Abstract

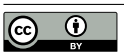
In shared immersive media, expressive avatars can provide social context, emotional feedback, or companionship. However, poorly aligned avatar reactions may also draw attention away from the primary content. We investigate this in a VR eye-tracking pilot in which participants watched 30-second emotional video clips alongside a co-present MetaHuman avatar expressing either matching or mismatching emotions. Across 28 participants, mismatching avatar reactions were associated with higher avatar dwell time than matching reactions, longer individual glances, and stronger early inter-subject synchrony of avatar-directed gaze. These observations suggest that avatar reaction mismatch is a promising factor in VR attention design, with implications for virtual cinemas, remote events, training simulations, and avatar-mediated communication. The study used a fixed presentation order, with matching trials preceding mismatching trials. This limitation is addressed in the discussion and motivates future counterbalanced work.

CCS Concepts

• **Human-centered computing** → *Empirical studies in HCI*; **Virtual reality**; • **Computing methodologies** → **Perception**; *Animation.*

Keywords

Avatar, Virtual Reality, Gaze, Facial Animation, Co-Viewing



This work is licensed under a Creative Commons Attribution 4.0 International License.
SAP '26, Rennes, France
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2799-3/26/08
https://doi.org/10.1145/3821409.3833897

ACM Reference Format:
Joshua Smallberg, Sina Kitz, Jan Spindler, Bardia Ardakanian, Leif Van Holland, Nils Wandel, and Reinhard Klein. 2026. Congruency-Dependent Gaze Allocation to Avatar Facial Expressions during Shared Video Viewing in Immersive VR. In *ACM Symposium on Applied Perception 2026 (SAP '26), August 26–27, 2026, Rennes, France.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3821409.3833897

1 Introduction

Shared immersive media increasingly involves expressive virtual humans. In virtual cinemas, remote events, training simulations and avatar-mediated communication, co-present avatars may provide companionship, emotional feedback, or social context. Yet these avatars may also compete with the primary content for visual attention, especially when their reactions do not align with the presented media.

This raises an important design question for social VR. When do avatar reactions remain peripheral context and when do they become attentional targets? Prior work shows that facial expressions are interpreted in relation to emotional context [Righart and de Gelder 2008] and that inter-subject correlation can reveal stimulus-locked attention during naturalistic viewing [Hasson et al. 2004; Nastase et al. 2019]. However, less is known about how attention is distributed when emotional video content and a co-present expressive avatar are viewed simultaneously in VR.

We examine this question in an immersive VR co-viewing setup. Our contribution is a compact empirical characterization of avatar reaction mismatch as a potential attentional design factor. We report effects on avatar dwell, glance duration, and inter-subject synchrony.

2 Study Design

Participants and VR setup. Twenty-eight participants watched eight 30-second trials in a Meta Quest Pro headset. The headset

rendered at 1800 × 1920 pixels per eye at 90 Hz, and its integrated eye tracker recorded gaze at 90 Hz with an approximate spatial accuracy of 1.1 degrees. Four videos were selected from the VEATIC dataset [Ren et al. 2024] to represent happy, angry, sad, and neutral affect. The experiment took place in a virtual living-room scene containing a wall-mounted video screen and a co-present MetaHuman avatar. The screen was placed to the participant's right and the avatar to the left at approximately 1.0 – 1.2 m viewing distance. Both were visible without head movement, but fixating the avatar required shifting gaze away from the video.

Stimuli and animation control. Avatar facial animations were generated using FEELS (Fine-grained Expression Editing via Latent-space Semantic-classification), a test-time guidance method for temporally controlled facial animation, and then retargeted from FLAME [Li et al. 2017] parameters to a MetaHuman character in Unreal Engine 5. Guidance parameters were fixed across trials and each avatar emotion appeared once in a matching condition and once in a mismatching condition. The avatar was rendered as a non-interactive expressive co-viewer, avoiding additional confounds from dialogue, gesture, or contingent social behavior. Additional analyses of blendshape motion energy are reported in the supplementary material.

Procedure and measures. The eight trial conditions were H+H, A+A, N+N, S+S, H+N, A+H, N+S, and S+A, where the first letter denotes video emotion and the second avatar emotion. H, A, N, and S denote happy, angry, neutral, and sad. Participants were told that the task involved co-viewing videos with an avatar in a living-room setting, but received no further instructions about the avatar's emotional behavior. The first four trials were matching conditions, followed by four mismatching conditions using the same video set.

Participants were instructed to watch the videos naturally and were not told to fixate either the screen or the avatar.

The primary dependent variable was avatar dwell time, defined as the proportion of gaze samples falling on the avatar. We also analyzed glance rate and mean glance duration to distinguish whether dwell differences reflected more frequent avatar glances or longer individual glances. Finally, we computed leave-one-out inter-subject correlation (LOO-ISC) of avatar gaze using 1-second bins over the 30-second trial. Early-window ISC was computed over the first four seconds after trial onset.

3 Results

Mismatching avatar reactions were associated with more avatar-directed gaze than matching reactions. Mean avatar dwell increased from 22.8% to 33.9%, a paired difference of 11.1 percentage points, $t(27) = 4.50$, $p < .001$, $d_z = .85$ (Figure 1, top).

Glance rate was nearly identical across condition types, but individual glances were longer under mismatch ($M = 1.71$ s) than match ($M = 1.14$ s), $W = 53$, $p < .001$. The mismatch effect was not limited to greater overall dwell time. Avatar gaze was also more synchronized across participants in mismatching trials, indicating that independent observers tended to look toward the avatar at similar moments in time. A per-participant paired test, comparing each observer's mean LOO-ISC (Fisher-z scale) across the four congruent versus the four incongruent conditions, confirmed that

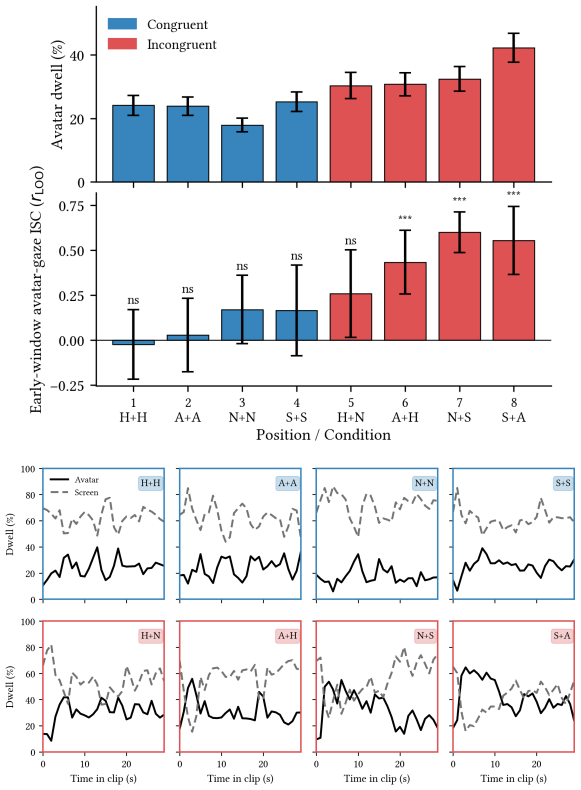


Figure 1: Dwell and synchronization by condition. Top: avatar dwell (±1 SE) and early-window ISC (r_{LOO}; 95% CI). Significance markers reflect Holm–Bonferroni-corrected p-values across $m = 8$ simultaneous one-sample t-tests against zero ($^{***}p_{Holm} < .001$; ns = not significant); bottom: avatar/screen dwell time courses.

avatar-gaze synchrony was greater under incongruence (congruent $r = 0.019$, incongruent $r = 0.175$; $t(27) = 6.48$, $p < .001$, $d_z = 1.23$).

During the first four seconds after trial onset, three out of four mismatching conditions showed significant LOO-ISC after Holm correction, whereas no matching condition did (Figure 1, top). The corresponding avatar/screen time courses are shown in Figure 1, bottom. Thus, mismatching avatar reactions elicited not only more gaze to the avatar, but a temporally precise shared gaze response across observers.

4 Discussion

This work characterizes avatar reaction mismatch as a potential attentional design factor in shared immersive media. In our VR co-viewing pilot, mismatching avatar reactions were associated with increased avatar dwell, longer individual glances and stronger early inter-subject synchrony of avatar-directed gaze. These observations suggest that a digital avatar can become an attentional target when its facial expression differs from the affective context of co-viewed media.

A key observation of our study is the temporal synchrony of gaze. Beyond the mean dwell increase, mismatching trials showed

stronger alignment in when participants looked toward the avatar. This suggests that incongruent avatar reactions can create stimulus-locked attentional events, where viewers do not merely look more at the avatar overall, but tend to look toward it at similar points in time. Such temporal precision could be useful for VR design because it indicates when an avatar reaction becomes attentionally salient enough to compete with the primary media content.

The main limitation of the study is the fixed stimulus order: matching trials preceded mismatching trials. As a result, the data cannot disentangle affective mismatch from block order, repeated video exposure, or task learning, and should be interpreted as an empirical observation. These results motivate future work that counterbalances trial order while testing whether avatar reaction mismatch can be used deliberately to guide or avoid attention shifts in shared immersive media.

Acknowledgments

This work was supported by the European Regional Development Fund (ERDF) and the State of North Rhine-Westphalia as part of the operational program EFRE/JTF-Programm NRW 2021–2027. The project, titled "Gen-AIvatar", was funded under the NEXT.IN.NRW competition with the grant agreement No. EFRE-20801085.

The work has also been funded by the Ministry of Culture and Science North Rhine-Westphalia under grant number PB22-063A (InVirtuo 4.0: Experimental Research in Virtual Environments), and by the state of North Rhine-Westphalia as part of the Excellency Start-up Center.NRW (U-BO-GROW) under grant number 03ESCNW18B.

References

Uri Hasson, Yuval Nir, Ifat Levy, Galit Fuhrmann, and Rafael Malach. 2004. Intersubject Synchronization of Cortical Activity During Natural Vision. *Science (New York, N.Y.)* 303 (04 2004), 1634–40. doi:10.1126/science.1089506

Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.* 36, 6, Article 194 (Nov. 2017), 17 pages. doi:10.1145/3130800.3130813

Samuel A. Nastase, Valeria Gazzola, Uri Hasson, and Christian Keysers. 2019. Measuring shared responses across subjects using intersubject correlation. *Social Cognitive and Affective Neuroscience* 14, 6 (08 2019), 667–685. doi:10.1093/scan/nsz037

Zhihang Ren, Jefferson Ortega, Yifan Wang, Zhimin Chen, Yunhui Guo, Stella X. Yu, and David Whitney. 2024. VEATIC: Video-based Emotion and Affect Tracking in Context Dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. IEEE Computer Society, Los Alamitos, CA, USA, 4467–4477. doi:10.1109/WACV57701.2024.00441

Ruthger Righart and Beatrice de Gelder. 2008. Recognition of facial expressions is influenced by emotional scene gist. *Cognitive, affective & behavioral neuroscience* 8 (10 2008), 264–72. doi:10.3758/CABN.8.3.264