GazeFlow: Diverse Driver Gaze Synthesis via Conditional Flow Matching

Yonghyun Kim*
Electronics and Telecommunications Research Institute (ETRI)
Daejeon, Republic of Korea
yhkeen@etri.re.kr

Woojin Kim
Electronics and Telecommunications Research Institute (ETRI)
Daejeon, Republic of Korea
wjinkim@etri.re.kr

Hyunsuk Kim
Electronics and Telecommunications Research Institute (ETRI)
Daejeon, Republic of Korea
hyunsuk@etri.re.kr

Eun Hye Jang
Electronics and Telecommunications Research Institute (ETRI)
Daejeon, Republic of Korea
cleta4u@etri.re.kr

Daesub Yoon
Electronics and Telecommunications Research Institute (ETRI)
Daejeon, Republic of Korea
eyetracker@etri.re.kr

Abstract

Modeling diverse human gaze behavior is a fundamental challenge in behavioral simulation and character animation for interactive graphics and digital humans. In safety-critical applications such as Driver Monitoring Systems (DMS) under the European New Car Assessment Programme (Euro NCAP) 2026, existing synthesis approaches produce deterministic outputs that lack the stochastic diversity of real drivers. To overcome this limitation, we cast gaze-motion synthesis as a stochastic coverage problem anchored to a measured human intra-condition diversity baseline of $D_{\text{human}} = 1.94$, defined as mean pairwise trajectory distance under identical conditioning, computed from 59,405 condition-matched subject pairs across 689 scenario groups. A conditional flow matching model over 5D gaze-head trajectories recovers 93% of this baseline at our human-calibrated operating point ($\sigma = 0.8$). A strong conditional variational autoencoder (CVAE) baseline, under matched conditioning, saturates at 41% of human diversity regardless of latent scale. The sampling noise scale σ further enables continuous fidelity–diversity trade-off control: at $\sigma = 0.6$ the model achieves absolute position error (APE) within 7% of the CVAE while covering 64% of human variability.

CCS Concepts

• **Computing methodologies → Motion processing**.

ACM Reference Format:
Yonghyun Kim, Woojin Kim, Hyunsuk Kim, Eun Hye Jang, and Daesub Yoon. 2026. GazeFlow: Diverse Driver Gaze Synthesis via Conditional Flow Matching. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Posters (SIGGRAPH Posters '26), July 19–23, 2026, Los Angeles, CA, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3799825.3818715

*Corresponding author.

SIGGRAPH Posters '26, Los Angeles, CA, USA

ACM ISBN 979-8-4007-2548-7/26/07
https://doi.org/10.1145/3799825.3818715

1 Introduction

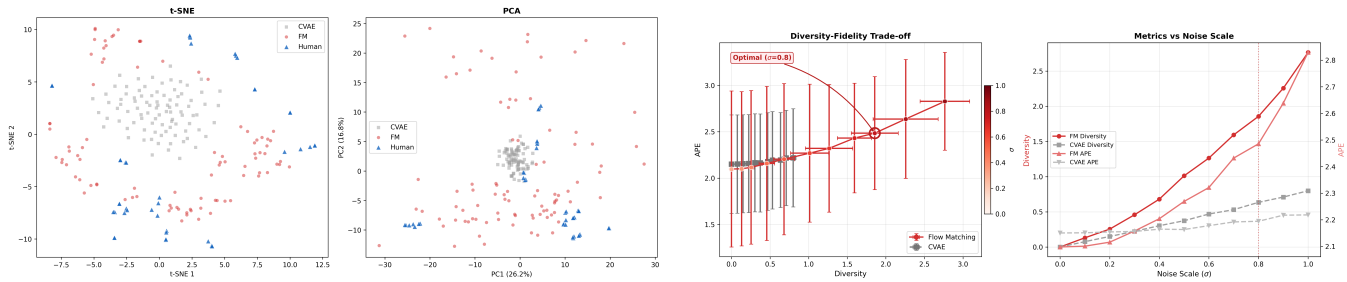
Diverse and realistic gaze behavior is essential for attention-driven character controllers and digital human simulation in interactive graphics and behavioral animation. In safety-critical applications such as Driver Monitoring Systems (DMS) under the Euro NCAP 2026 Safe Driving protocol [European New Car Assessment Programme 2025], which evaluates driver distraction across Long, Short, Phone-Use categories, movement types (Owl, Lizard, Body Lean), and diverse demographics, generating such behavior at scale remains challenging. Covering this combinatorial space with human-driven trials alone is impractical, motivating synthetic gaze-motion data as a scalable evaluation substrate.

Prior synthesis methods produce deterministic keyframe sequences from areas-of-interest (AOI) instructions. Real drivers, however, show substantial variability in saccade velocity, head-rotation strategy, and gaze-transition dynamics [Shiferaw et al. 2018]. A DMS validated only on mode-collapsed data may fail on the safety-critical tails.

While diffusion-based approaches rely on complex iterative denoising schedules, conditional flow matching directly learns the probability flow between a simple prior and the data distribution. This straight-path formulation allows more direct control over the fidelity–diversity trade-off by scaling the prior variance σ, enabling us to precisely calibrate the model's stochastic coverage to match the empirical human baseline.

We formulate gaze-motion synthesis as a stochastic coverage problem anchored to empirical human variability and make three contributions:

(1) We establish a human intra-condition diversity baseline $D_{\text{human}} = 1.94 \pm 0.88$ from 59,405 condition-matched subject pairs across 689 scenario groups.
(2) We present a conditional flow matching model [Lipman et al. 2023] that reaches 93% of this baseline at a human-calibrated



(a) (b)

Figure 1: Structured stochasticity analysis. (a) PCA and t-SNE projections of generated trajectory distributions at $\sigma = 0.8$: flow matching produces conditionally consistent yet multi-modal trajectories that match the full spread of human data, while the CVAE collapses near the conditional mean. (b) Fidelity–diversity Pareto trade-off as a function of the noise scale $\sigma \in [0, 1]$; flow matching reaches the human baseline at the calibrated point $\sigma = 0.8$.

operating point, with continuous trade-off control via the sampling noise scale σ.

(3) We show that a strong CVAE baseline with identical conditioning empirically saturates at 41% of human diversity across the latent-scale sweep.

2 Our Approach

2.1 Data Collection

We collected data from 109 licensed drivers (ages 20–50, gender-balanced) in a CARLA-based Level-2 simulator. A Smart Eye Pro 8.0 tracker (~120 Hz) captured head pose and gaze direction. Seven protocols combined visual and physical distractions, yielding 581 sessions (total 31.5 h) with gaze targets in seven AOIs [Chang et al. 2025].

2.2 Gaze Trajectory Representation

Each frame is a 5D vector $\mathbf{x}_t = (h_{\text{yaw}}, h_{\text{pitch}}, h_{\text{roll}}, g_{\text{yaw}}, g_{\text{pitch}}) \in \mathbb{R}^5$ (radians, 10 Hz). Recordings are segmented into 15 s windows (150 frames) with 5 s stride, producing 49,590 augmented segments after time-jitter and Gaussian noise. The split is 80/10/10 by subject.

2.3 Conditional Flow Matching

Each segment is conditioned on a 20D vector $\mathbf{c}$ (distraction category, movement type, AOI one-hot, duration, gender, age group). The model learns the velocity field of the linear interpolant $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$ by regressing $v_\theta(\mathbf{x}_t, t, \mathbf{c})$ onto the ground-truth displacement $(\mathbf{x}_1 - \mathbf{x}_0)$ under a squared-error objective. The backbone is a 4-layer transformer encoder ($d = 256$, 4 heads, 3.5M parameters). Inference uses 50-step Euler integration from $\mathbf{x}_0 \sim \mathcal{N}(0, \sigma^2\mathbf{I})$. A CVAE (1.1M parameters) with identical conditioning serves as the baseline.

3 Results

Evaluation uses 50 condition-stratified held-out scenario groups (30 samples each). Metrics follow [Tevet et al. 2023] in standardized feature space. Diversity is mean pairwise L_2 distance under identical conditioning.

3.1 Human Baseline and Operating Point

We extracted 59,405 condition-matched subject pairs to obtain $D_{\text{human}} = 1.94 \pm 0.88$. Sweeping σ identifies $\sigma = 0.8$ as the primary human-calibrated operating point (diversity 1.81, within 7% of baseline). $\sigma = 0.6$ offers a fidelity-priority alternative (APE within 7% of CVAE while retaining 64% coverage).

3.2 Main Results

Table 1 summarizes trajectory-level evaluation. Under matched conditioning, the CVAE saturates at 41% coverage: scaling its prior lifts diversity only up to saturation while leaving APE essentially unchanged, a pattern consistent with posterior over-regularization under strong conditioning rather than under-capacity, given that the CVAE already attains the lowest APE among all models. Flow matching at $\sigma = 0.8$ reaches 93% of the human baseline, and the coverage gain from 41% to 93% is the practically meaningful result, with $\sigma = 0.6$ offering near-APE parity to the CVAE while still covering 64% of variability. Figure 1(a) and Figure 1(b) confirm that flow matching strictly dominates the CVAE Pareto front. Removing demographic conditioning (20D $\longrightarrow$ 15D) worsens APE and diversity simultaneously, and between-group differences are themselves statistically significant (e.g., head-yaw velocity: Male 20s 77.3 ± 120.0 vs. Male 40s 62.1 ± 109.1 deg/s, $t(6736) = 5.32$, $p < .001$), indicating that demographic factors carry real conditioning signal. Finally, APE ordering across distraction categories mirrors human diversity, confirming that APE alone is variance-confounded and supporting coverage-based evaluation as the primary axis.

4 Conclusion and Future Work

We recast driver gaze-motion synthesis as a stochastic coverage problem anchored to a measured human diversity baseline, turning "more diverse" into a falsifiable claim. Conditional flow matching closes this coverage gap where a strong CVAE does not, and the sampling noise scale σ exposes a continuous fidelity–diversity trade-off within a single model. Beyond DMS, this lightweight real-time 5D gaze generator serves as a building block for attention-driven character controllers in interactive graphics and digital humans [Tevet et al. 2023]. Future work will incorporate eyelid and pupil dynamics and explore continuous rotation parameterizations [Zhou et al. 2019] at larger scale.

Acknowledgments

This work is supported by Korea Evaluation Institute of Industrial Technology (KEIT) grant funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea) (No. 20024892; RS-2024-00403108).

References

Mi Chang, Eun Hye Jang, Woojin Kim, and Daesub Yoon. 2025. Gaze Entropy and Driver Safety: Understanding Cognitive Failure and Situational Response Before Take-over. In *ACM SIGGRAPH 2025 Posters (SIGGRAPH Posters '25)*. ACM. doi:10.1145/3721250.3742974

European New Car Assessment Programme. 2025. *Euro NCAP Assessment Protocol — Safe Driving: Driver Engagement*. Technical Report. Euro NCAP. https://cdn.euroncap.com/cars/assets/euro_ncap_protocol_safe_driving_driver_engagement_v11_a30e874152.pdf Version 1.1.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2023. Flow Matching for Generative Modeling. In *International Conference on Learning Representations (ICLR)*. doi:10.48550/arXiv.2210.02747

Brook A. Shiferaw, Luke A. Downey, Justine Westlake, et al. 2018. Stationary gaze entropy predicts lane departure events in sleep-deprived drivers. *Scientific Reports* 8, 1 (2018), 2220. doi:10.1038/s41598-018-20588-7

Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-Or, and Amit H. Bermano. 2023. Human Motion Diffusion Model. In *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=SJ1kSyO2jwu

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. 2019. On the Continuity of Rotation Representations in Neural Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5745–5753. doi:10.1109/CVPR.2019.00589

Table 1: Trajectory-level evaluation sorted by human coverage. Bold = our main result ($\sigma = 0.8$). Values are mean ± SD.

Model	APE ↓	DTW ↓	Diversity	% Human
Human (reference)	—	—	1.938 ± 0.884	100%
FM ($\sigma = 0.8$, human-calib.)	2.643 ± 0.797	2.637 ± 0.804	**1.807 ± 0.353**	**93%**
FM ($\sigma = 0.6$, alternative)	2.475 ± 0.785	2.509 ± 0.732	1.241 ± 0.250	64%
CVAE ($\sigma = 1.0$, max coverage)	2.359 ± 0.531	2.344 ± 0.533	0.804 ± 0.033	41%
CVAE ($\sigma = 0.6$, min APE)	**2.314 ± 0.529**	**2.330 ± 0.523**	0.460 ± 0.020	24%