GAZE WORLD MODEL

Anonymous authors
Paper under double-blind review

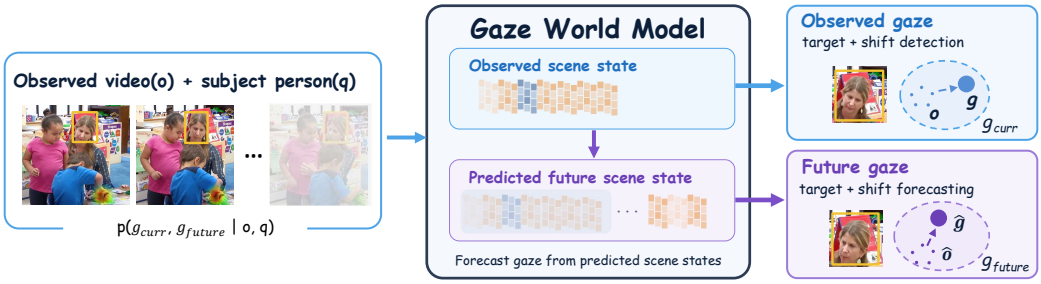

Figure 1: Overview of the Gaze World Model. Given the observation, the model estimates current gaze of the subject person and rolls forward latent scene dynamics to forecast future gaze targets and shift events beyond the observation window.

ABSTRACT

Where a person looks indicates their focus of attention, and how their gaze shifts between targets reveals how that attention responds to a changing scene. Existing gaze target estimation methods largely operate frame by frame and do not model these shifts. Gaze shifts are often driven by dynamic changes in the scene, which is the kind of information world models are designed to encode. This relationship is subtle, however, since an event far across the frame may produce only a slight movement of the person's head or eyes. We therefore introduce the Gaze World Model (GWM), which co-trains latent scene prediction with person-conditioned gaze and shift supervision on a frozen V-JEPA 2.1 encoder, sensitizing the model to how scene changes relate to a specific person's attention. Given a video clip and the person's head box, GWM estimates gaze locations and detects gaze shifts in the observed clip, and forecasts future gaze and shifts from predicted scene features. To train and evaluate these tasks, we augment VideoAttentionTarget and ChildPlay with event-level gaze-shift annotations. GWM achieves state-of-the-art performance on gaze target estimation and shift detection on these benchmarks, and outperforms the compared baselines on future gaze localization and shift forecasting.

1 INTRODUCTION

Social gaze, the process of attending to a social partner and inferring, responding to, or directing their attention, is a key building block of social communication (Tomasello & Farrar, 1986). In developmental contexts, children use their gaze to communicate well before they understand spoken language. For example, during play a child may first look at an adult's face and then shift their gaze to a toy that the adult is holding (Land, 2006; Frischen et al., 2007). The child's gaze location indicates their current focus of attention, and the shift from the adult to the toy could be part of a requesting behavior, where the child communicates their interest. Thus, while gaze locations identify the visual inputs a child is receiving, the timing of the child's gaze shifts in the social context, which signal their intent and engagement, convey information that can't be extracted from the gaze locations alone.

Significantly, most prior works on gaze target estimation (GTE) have focused on localizing gaze within a single image (Recasens et al., 2015; Lian et al., 2018; Chong et al., 2018; Ryan et al., 2025; Ye et al., 2026). While a few video-based methods added temporal context, this consisted primarily of temporal smoothing over independently identified gaze locations (Chong et al., 2020;

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.

Tafasca et al., 2023). In these prior works, gaze is estimated as a point in each frame, and transitions between gaze targets are not explicitly modeled or predicted. The goal of this paper is develop a model which is capable of *explicitly predicting temporal shifts in gaze*, along with the evidence for those shifts in the form of changing gaze locations over time.

A natural approach is to treat gaze-shift detection as an action recognition problem, and learn the temporal signature of a shift, such as a head turn (Wang et al., 2025). But a gaze shift is always accompanied by a change in the gaze target, and accurate inference therefore requires an understanding of the structure of the scene as well as the person. Since gaze shifts are often driven by the events in a person's environment, such as the introduction of a novel object, an effective model for gaze shift detection must be attuned to the manner in which scenes change over time. Fortunately, recent works in world modeling have introduced a powerful approach to encode the current state of an environment and predict how it evolves over time (Ha & Schmidhuber, 2018; Hafner et al., 2025).

Pretrained world-model encoders such as V-JEPA (Assran et al., 2025; Mur-Labadia et al., 2026) capture how a scene evolves within the observed video, but they are not trained to predict beyond it, nor to model how a particular person might respond to these changes. The relationship between a scene change and the resulting gaze shift can be subtle, as a person's face or eyes may change only slightly in response to distal event, whereas world models are primarily effective at capturing causal relationships that are readily visible in the video. In order to overcome this challenge, we develop a procedure for co-training latent scene prediction with person-conditioned gaze and shift supervision on top of a frozen V-JEPA encoder. This approach achieves state-of-the-art results for both gaze shift detection and gaze location estimation. We are the first to explicitly detect and model gaze shifts and leverage the world modeling approach for gaze understanding.

Building on this insight, we propose the Gaze World Model (GWM; Figure 1). Given an observed clip and a head box for the person of interest, GWM encodes the observed scene with a frozen V-JEPA 2.1 encoder (Mur-Labadia et al., 2026), conditions the scene features on the person's head through scene–head interaction, and aggregates them over time to estimate gaze and detect shifts. A co-trained world predictor predicts future scene features, from which gaze decoders forecast future gaze and shifts. At inference, GWM requires only the observed video and the head box. Our contributions are as follows:

- We formulate gaze modeling for a specified person as a world-modeling problem, covering gaze point estimation, gaze-shift detection, and forecasting over observed and future intervals.
- We propose GWM, which co-trains latent world prediction with person-conditioned gaze and shift supervision on a frozen V-JEPA 2.1 encoder, sensitizing the model to the subtle relationship between scene changes and gaze shifts.
- We introduce a gaze-shift dataset built on VideoAttentionTarget and ChildPlay, with event-level gaze-shift annotations over 4 hours of person-level gaze tracks, for training and evaluating gaze-shift detection and forecasting.

2 Related Work

Human Gaze Modeling. Gaze following predicts a spatial target distribution from a person's appearance and head orientation (Recasens et al., 2015; Lian et al., 2018; Chong et al., 2018). Video methods add temporal context and out-of-frame prediction (Chong et al., 2020; Tafasca et al., 2023), while recent spatial models use frozen foundation encoders and stronger decoders (Ryan et al., 2025; Ye et al., 2026). Subject conditioning identifies whose gaze to predict, through head boxes or textual and multimodal cues (Cao et al., 2026; She et al., 2026; Miao et al., 2026). Gaze-event analysis and anticipation address a different temporal question: classical methods identify fixations and saccades from ocular signals (Salvucci & Goldberg, 2000; Holmqvist et al., 2011; Hessels et al., 2025), and anticipation methods predict future gaze or behavior (Zhang et al., 2017; Li et al., 2018; 2026; Furnari & Farinella, 2021; Kuang & Agarwal, 2026). We study target-transition events for a specified person, with detection in observed video and forecasting beyond it. Appendix A provides the extended discussion.

World Models and Behavioral Prediction. World models learn transition dynamics that support prediction and planning by rolling the current state forward under possible actions (Ha & Schmidhuber, 2018; Hafner et al., 2025; Ding et al., 2025). Behavioral world modeling shifts the emphasis

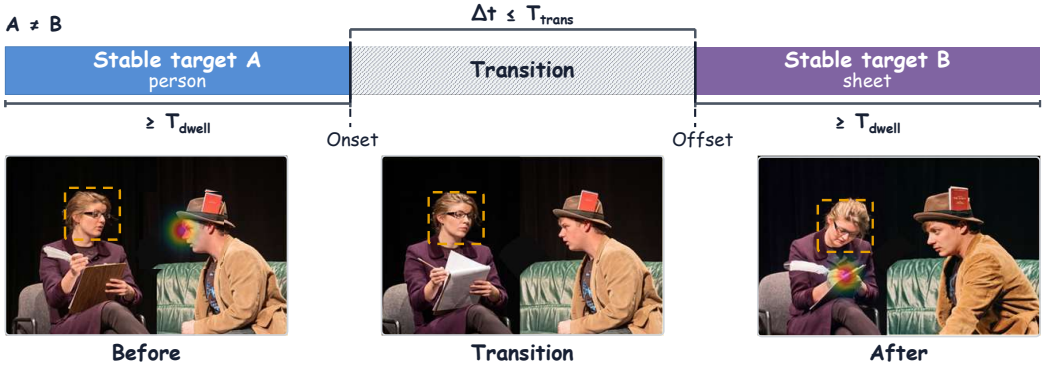

Figure 2: Definition of a gaze-shift event. Per-frame target labels of the specified person (dashed box) show a stable fixation on A, a transition interval, and a stable fixation on B.

from predicting scene appearance alone to predicting how agents act and how their interactions change the world (Shen et al., 2026). Existing approaches learn structured human-centric actions and affordances from human videos (Mendonca et al., 2023), or jointly predict an egocentric agent's future action and world state (Chen et al., 2025); related gaze–action models likewise use attention as a cue to future human behavior (Huang et al., 2015; Kuang & Agarwal, 2026). These models primarily address embodied control or camera-wearer actions. We model a specified person's semantic gaze target in third-person video as a compact behavioral state: target changes define observed events, and world prediction supports future gaze and shift forecasting.

3 Gaze World Model

The Gaze World Model (GWM) estimates and forecasts gaze for a specified person. We distinguish annotated events, benchmark labels, and model outputs, then describe how subject conditioning, temporal aggregation, and latent prediction support these tasks.

3.1 Task Definition

In Figure 2, the specified person shifts gaze from another person (A) to a sheet (B), illustrating a human-to-other transition. A gaze shift consists of stable fixations on two different targets ($A \neq B$), each lasting at least T_{dwell}, separated by a transition interval of at most T_{trans} (Appendix Table 8). We define three person-involving transition types as positive: human-to-human between distinct people, human-to-other, and other-to-human. Here, `other` includes objects, background regions, and out-of-frame targets. Changes between body parts of the same person or between `other` targets are negative. Alongside gaze-point estimates, detection assigns each observed window one of the three transition labels or no in-scope shift. Forecasting predicts whether a transition of any of the same three types begins in the future window. Neither task predicts exact event boundaries or target instance identifiers. Both tasks use the cleaned annotations produced by our data pipeline. Appendix Table 7 summarizes the terminology and provides the window-label assignment rule.

3.2 Model Architecture

Figure 3 summarizes one architecture with a shared frozen visual encoder, scene–head interaction (SIM), a temporal trunk, and a world-prediction branch. Three groups of task heads provide gaze heatmaps and in/out predictions, gaze-shift detection, and gaze-shift forecasting.

Scene Representation Encoder. A frozen V-JEPA 2.1 encoder (Mur-Labadia et al., 2026) encodes the observed scene frames. In the main forecasting setting, GWM observes 16 frames and forecasts gaze over the next 8 frames at 8 FPS. Each tubelet spans two frames, giving $T_o = 8$ observed and $T_f = 4$ future positions, and each scene tubelet yields $P = 32 \times 32 = 1024$ patch features of width C. Future frames are never encoded on the prediction path; they are used only to form latent targets during forecasting training.

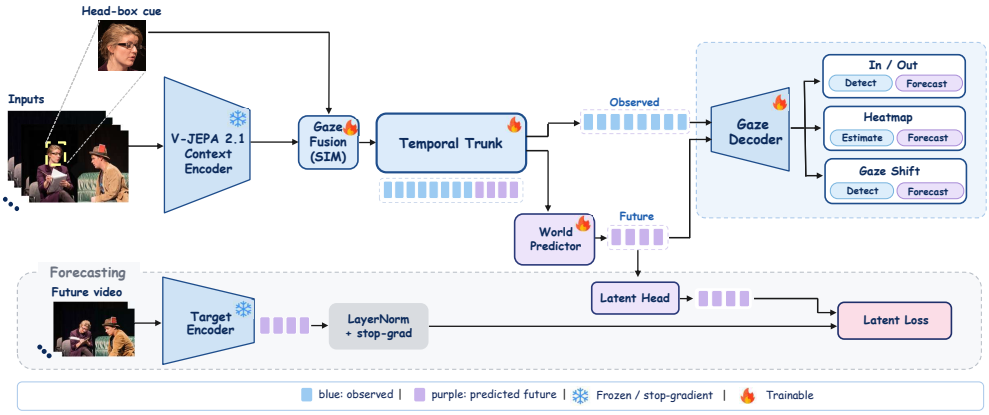

Figure 3: Architecture of GWM. The frozen V-JEPA 2.1 encoder encodes the observed clip, and SIM conditions the scene features on the head box. The temporal trunk processes observed and future frames tokens, and the world predictor refines the future positions. A shared gaze decoder reads the observed positions for estimation and detection, and the predicted future positions for forecasting.

Subject-Conditioned Fusion. The scene–head interaction module (SIM) conditions the scene features on the specified person. It pairs the scene stream with a head stream encoded from the person's head crops (Appendix B), and linear projections map both streams to width $D = 256$. Each stream is prefixed with $R = 5$ tokens, an in/out token at position zero and four register tokens, so a scene tubelet has $N = P + R = 1029$ tokens. The head box, averaged over the two frames of a tubelet, places the head tokens in the scene by assigning them RoPE coordinates inside the box's scene-grid rectangle. After one self-attention layer per stream, five SIM layers each apply bidirectional scene–head cross-attention and stream-specific self-attention. SIM is pretrained and kept frozen during forecasting training.

Temporal Trunk. The trunk aggregates person-conditioned scene features across tubelets using joint self-attention with three-dimensional RoPE coordinates (t, y, x) and SwiGLU MLPs. Each scene position retains its N tokens. For forecasting, a normalized linear projection of the raw observed encoder features is added to the corresponding SIM patch tokens, with zero additions at prefix positions. Four future query bundles are initialized by copying the last observed bundle (index 7) and adding a learned future embedding. The two trunk blocks process all $T_o + T_f = 12$ positions. Observed queries attend only to observed positions; future queries attend to both observed and future positions. Thus, attention masking controls access among token positions, while the future queries themselves contain no future pixels. The head-token stream bypasses both the trunk and the world branch.

World Predictor. The world branch applies two additional RoPE transformer blocks to all 12 trunk positions with the same attention mask. Observed gaze predictions read the first eight trunk outputs, while future gaze predictions read the last four world-branch outputs. Future positions originate from the copied observation and future embedding.

In the forecasting stage, we supervise future patch representations by latent prediction (Assran et al., 2023; Bardes et al., 2024). During training only, the same frozen encoder separately processes the eight real future scene frames without gradient tracking. Its patch features are layer-normalized and detached to form targets $z_{t,p}$, where $t \in \{8, \ldots, 11\}$ and $p \in \{1, \ldots, P\}$. Let $e_{s,p}$ denote the raw observed encoder features and $u_{t,p}$ the world-branch patch outputs. The latent prediction is a residual on the normalized observation mean,

$$\hat{z}_{t,p} = A\ \mathrm{LN}(u_{t,p}) + b + \mathrm{LN}\left(\frac{1}{T_o}\sum_{s=0}^{T_o-1} e_{s,p}\right),$$

where $A : \mathbb{R}^{256} \to \mathbb{R}^C$ and b are zero-initialized. The motion-weighted objective is

$$\mathcal{L}_{\mathrm{lat}} = \frac{1}{|\mathcal{M}|C} \sum_{(t,p)\in\mathcal{M}} w_{t,p}\, \|\hat{z}_{t,p} - z_{t,p}\|_1. \tag{1}$$

The weights $w_{t,p}$ emphasize patches with larger temporal feature changes and are detached from gradient computation. Here $\mathcal{M}$ contains the 4×1024 future patch positions per sample, excluding all prefix tokens, and the L1 norm sums over the C feature channels. The loss is also averaged over the batch and contributes $\lambda_{\text{lat}}\mathcal{L}_{\text{lat}}$ with $\lambda_{\text{lat}} = 0.1$. Targets come from the shared frozen encoder; there is no EMA encoder or target predictor. Separate observed/future encoding keeps future pixels outside the prediction path, including during training.

Task Decoders. Let $G_{t'}$ denote the scene tokens selected from the trunk for observed positions or from the world branch for future positions, and let $H_{t'}$ denote the observed SIM head tokens.

- **Gaze heatmap and in/out heads.** A shared heatmap head upsamples the scene patch grid with a transposed convolution and a 1×1 convolution to obtain 64×64 heatmap logits $\hat{M}_{t'}$. For observed positions, the in/out head concatenates the position-zero scene and head tokens and applies an MLP to obtain the logit $\hat{v}_{t'}$; its sigmoid gives the in-frame probability. Future positions reuse the last observed head in/out token, $H_{7,0}$. Both $\hat{M}$ and $\hat{v}$ are logits; training uses binary cross-entropy with logits on valid labels. Each tubelet prediction is repeated over its two constituent frames to align with frame-level supervision.
- **Gaze-shift detection head.** Each observed tubelet is summarized by $x_{t'} = W_s[\bar{g}_{t'}; q_{t'}; m_{t'}; h_{t'}] + b_s$. Here $\bar{g}_{t'} = R^{-1}\sum_{r=0}^{R-1} G_{t',r}$ averages all five scene prefix tokens, $q_{t'}$ pools scene patches with heatmap-derived attention, $m_{t'}$ is an MLP embedding of seven scalars (soft-argmax x, y, heatmap peak, in/out probability, $\Delta x, \Delta y$, and displacement magnitude), and $h_{t'}$ averages all head-stream tokens. A transformer layer aggregates the eight summary tokens with temporal embeddings and a learned class token, followed by a linear classifier with $1 + 3$ logits for the negative class and the three positive transition types.
- **Gaze-shift forecasting head.** A separate head uses the same four-part summary design and temporal aggregation, but has two output logits for an in-scope shift versus no in-scope shift in the future interval. It reads all 12 tubelets: eight observed trunk outputs and four future world-branch outputs. Observed pose summaries use the head-token mean; future summaries use a learned placeholder vector. The forecasting head is trained with two-class cross-entropy.

4 GAZE-SHIFT DATASET

Frame-level gaze coordinates are insufficient to supervise gaze shifts: they do not indicate whether neighboring coordinates refer to the same target instance, where a transition begins and ends, or which targets it connects. Exhaustively annotating these events in unconstrained video is costly and prone to disagreement between annotators. We therefore convert the frame-level annotations in VideoAttentionTarget (VAT) (Chong et al., 2020) and ChildPlay (Tafasca et al., 2023) into event-level transitions and window-level training labels, using a semi-automated engine that proposes candidate transitions from spatial and temporal rules, verifies them with a multimodal model, and submits sampled candidates to iterative human verification. Appendix C specifies the engine in full. Human targets are distinguished by person identifiers propagated from the rule stage. Held-object keys encode an association with a person, while independent objects use semantic keys; neither representation guarantees persistent object-instance identity.

We assign window labels using the person-involving gaze transitions defined in Section 3.1. This scope captures changes in interaction partner, speaker, or social focus while distinguishing `human` endpoints from `other` targets. Transitions between body parts of the same person, such as head-to-hand (`same_person`), and transitions between `other` targets (`other2other`) do not constitute an in-scope shift, even when gaze visibly moves. A no in-scope shift label therefore indicates the absence of the three benchmark transition types, not necessarily a stationary gaze point. Cases with unresolved event status or required annotations are excluded, rather than treated as negatives (Appendix C).

The source videos run at 25 to 30 fps, and event construction uses the native frame rate of each clip. After annotation, both datasets are sampled at 8 fps and extracted with a sliding window of 16 frames, so each model input covers 2 seconds. We split the data by source video. The annotation set contains 6,204 confirmed transitions, and training uses 15,086 windows.

5 EXPERIMENTS

We evaluate GWM on the two tasks defined in Section 3.1: detection within the observed clip and forecasting beyond it. Each task has two outputs for the specified person, the gaze points and the gaze shifts. We compare against three groups of methods: image-based GTE models that condition a frozen foundation encoder on the head box, a video gaze-anticipation baseline, and general multimodal models that receive the subject as a head box drawn on every frame. Section 5.2 first reports GTE and shift detection within the observed clip, then the temporal consistency of the predicted gaze trajectories, and finally forecasting beyond the clip. Section 5.3 examines the video-adaptation trunk, the latent prediction objective and its weight, the contribution of each loss term used for video training, and freezing the visual encoder; analyses of the class weighting in the shift detection loss and of the temporal sampling rate are also reported in Appendix E.

5.1 IMPLEMENTATION DETAILS

Baselines. We compare against Gaze-LLE (Ryan et al., 2025) and PaGE (Ye et al., 2026), each at two encoder sizes. For GTE, we evaluate the released models of Gaze-LLE and PaGE under the official dense video test protocol. For shift detection, we keep each image model frozen and train a two-layer temporal classifier on its per-frame features, in the same way as the shift detection head of GWM is trained on frozen upstream features. This setup contrasts two uses of time: GWM models temporal context before the gaze readout, whereas the image baselines encode each frame independently and aggregate time only in the added classifier. The image-model and GWM shift classifiers use the same training windows and test clips. For binary evaluation, a prediction in any of the three positive classes counts as shift; all others count as no shift. For forecasting, we compare against SAGE (Kuang & Agarwal, 2026), with its original detector, Sharingan, and with PaGE, the state of art image GTE model.

We also evaluate Qwen3.8 27B (Qwen Team, 2026) and Gemini 3.8 Flash (Gemini Team, 2026) zero-shot on all tasks, marking the specified subject with a red head box on every frame and prompting for probabilities, which serve as ranking scores for AP.

Training Schedule. GWM is a single model whose components are trained in three stages. In the image stage, we train subject conditioning and gaze readout on single frames, mainly from GazeFollow (Recasens et al., 2015), together with frames from the training splits of VideoAttentionTarget and ChildPlay, and MVGT (Miao et al., 2025). In the video detection stage, we first train the temporal trunk and the gaze heads on video with the heatmap, in/out, and latent prediction losses, and then train the gaze-shift detection head with all other components frozen. In the video forecasting stage, we initialize from the video model, train the world predictor together with the trunk and the gaze heads using the observed-gaze, future-gaze, and latent-prediction losses, and then train the binary forecasting head with all other components frozen; the input projections and SIM are also kept frozen in this stage.

By default, the V-JEPA 2.1 visual encoder is kept frozen in all stages; Table 6 examines unfreezing its last four blocks. Each input clip contains 16 observed frames sampled at 8 FPS (2 s), and in the main forecasting setting the model forecasts gaze for the following eight frames (1 s). The image stage is trained on $4\times$ NVIDIA H100 GPUs. The first step of the video detection stage trains for 20 epochs at an effective batch size of 32 with AdamW, a peak learning rate of 5×10^{-4}, a weight decay of 0.05, a linear warmup over 300 updates, and cosine decay to 1×10^{-7}. The shift detection head trains for 25 epochs at batch size 64 with the same schedule, a peak learning rate of 3×10^{-4}, and a warmup over 200 updates.

5.2 MAIN RESULTS

Table 1 compares GTE and gaze-shift detection within the observed clip. GWM (ViT-L) achieves lower gaze-point error on both datasets and higher scores on every shift metric than all external baselines. Across the full table, ViT-g leads in GTE and binary shift detection, while ViT-L best distinguishes transition types. Even our smaller ViT-B model outperforms PaGE ViT-H+, the strongest external shift-detection baseline, on every shift metric. These results support modeling temporal context before gaze readout: GWM encodes the observed clip over time, whereas the

Table 1: Observed gaze estimation and shift detection. GTE reports heatmap AUC, L2 in normalized image coordinates, and in/out AP as decimals; AUC is not reported (–) for the VLMs, which output gaze points without a probability heatmap. Event scores (%) are binary AP/F1 and four-class micro-AP/macro-F1; the three positive classes form binary shift. Bold marks the best values, including ties at the displayed precision.

Model	Params	Gaze-target estimation						Gaze-shift detection			
		VideoAttentionTarget			ChildPlay			Binary		Four-class	
		AUC↑	L2↓	AP↑	AUC↑	L2↓	AP↑	AP ↑	F1↑	AP ↑	F1↑
Specialized gaze models											
Gaze-LLE ViT-B	87M	0.945	0.089	0.885	0.953	0.108	0.998	68.46	62.24	51.31	49.59
Gaze-LLE ViT-L	304M	0.951	0.085	0.889	0.953	0.099	0.996	72.51	61.82	57.38	54.10
PaGE ViT-B	171M	0.971	0.063	0.946	0.966	0.082	0.998	76.91	74.07	63.08	60.16
PaGE ViT-H+	1.7B	0.971	0.065	0.943	0.964	0.081	0.997	78.38	74.42	66.32	65.67
Vision-language and generalist models											
Qwen3.8 27B	27B	–	0.155	0.789	–	0.134	0.981	62.74	52.88	44.35	49.47
Gemini 3.8 Flash	–	–	0.123	0.828	–	0.107	0.984	75.41	56.85	55.97	53.89
Ours											
GWM (ViT-B)	105M	0.966	0.061	0.942	0.968	0.073	0.994	85.34	78.50	72.45	71.80
GWM (ViT-L)	320M	0.973	0.057	0.948	0.974	0.066	0.997	88.65	82.17	**76.09**	**75.05**
GWM (ViT-g)	1B	**0.978**	**0.053**	**0.952**	**0.976**	**0.062**	**0.999**	**89.72**	**82.96**	75.91	74.87

Table 2: Temporal consistency of predicted gaze trajectories. Bold marks minima across all rows, including ties at the displayed precision.

Model	VideoAttentionTarget			ChildPlay		
	Jitter ↓	Flicker ↓	r ↓	Jitter ↓	Flicker ↓	r ↓
Gaze-LLE ViT-B	0.026	0.053	0.068	0.028	0.061	0.085
Gaze-LLE ViT-L	0.020	0.035	0.065	0.029	0.059	0.070
PaGE ViT-B	0.022	0.048	0.048	0.026	0.063	0.058
PaGE ViT-H+	0.015	0.030	0.054	0.020	0.043	0.060
− temporal trunk	0.019	0.037	0.050	0.020	0.048	0.057
GWM (ViT-B)	**0.012**	0.022	0.051	0.016	0.030	0.049
GWM (ViT-L)	**0.012**	0.020	**0.044**	**0.013**	**0.027**	0.048
GWM (ViT-g)	0.013	**0.019**	0.047	0.014	**0.027**	**0.045**

image baselines aggregate independently encoded frame features. Because the ViT-B model also leads in shift detection, this advantage does not depend on a larger backbone alone.

Table 2 compares predicted trajectories (Appendix D.1). ViT-L produces more stable fixations and a lower aligned movement residual than every external baseline on both datasets, supporting the value of temporal modeling beyond individual gaze-point estimates. ViT-g leads on VAT Flicker and ChildPlay movement residual, while ViT-L is best or tied on the other temporal metrics. Removing the trunk worsens both fixation stability and movement residual on both datasets, further supporting its role in modeling gaze dynamics. The aligned residual measures path accuracy after alignment, not response latency.

Table 3 shows that this advantage extends beyond the observed clip. Both ViT-L and ViT-g outperform all external baselines, including SAGE with PaGE, on future gaze localization at both reported horizons and on average, and on future-shift AP, F1, and recall. Across the full table, ViT-g achieves the lowest average gaze-point error on both datasets and leads shift forecasting. These results support predicting both where the specified person will look and whether a person-involving transition will begin.

5.3 Ablation studies

Unless otherwise stated, we use GWM with a frozen V-JEPA 2.1 ViT-L encoder for all ablation experiments on VideoAttentionTarget and ChildPlay. Additional results are provided in Appendix E.

Table 3: Gaze forecasting beyond the observation window. Avg. denotes the frame-wise mean L2 error over the future interval. Bold marks the best values, including ties at the displayed precision.

Gaze Model	Future gaze point (L2 ↓)						Future gaze shift		
	VideoAttentionTarget			ChildPlay			Binary		
	0.5 s	1.0 s	Avg.	0.5 s	1.0 s	Avg.	AP ↑	F1 ↑	Recall ↑
SAGE with Sharingan	0.137	0.167	0.142	0.162	0.181	0.166	56.51	65.65	78.30
SAGE with PaGE	0.116	0.148	0.123	0.143	0.165	0.145	66.23	66.67	71.40
Qwen3.8 27B	0.189	0.215	0.193	0.177	0.178	0.177	55.47	41.18	49.79
Gemini 3.8 Flash	0.168	0.210	0.180	0.152	0.167	0.155	62.00	44.20	53.00
Ours									
GWM (ViT-B)	0.117	0.145	0.120	0.125	0.146	0.128	71.30	71.20	84.20
GWM (ViT-L)	0.102	0.132	0.108	**0.123**	0.143	0.127	74.70	72.40	86.82
GWM (ViT-g)	**0.095**	**0.126**	**0.104**	**0.123**	**0.139**	**0.125**	**76.18**	**74.68**	**89.12**

Table 4: Ablation of temporal modeling for gaze detection and forecasting. Future L2 uses the same frame-wise average as Table 3. Without the temporal trunk, the world predictor has no input to roll forward, so forecasting results are not available (–).

Variant	Gaze-target estimation				Gaze-shift detection		Future gaze shift		
	VAT		ChildPlay		Four-class		Binary		
	L2↓ Observed	L2↓ Future	L2↓ Observed	L2↓ Future	AP↑	F1↑	AP↑	F1↑	Recall↑
−temporal trunk	0.069	–	0.076	–	71.25	69.97	–	–	–
−world predictor	0.068	0.118	0.073	0.135	73.92	73.71	71.69	70.64	84.18
$\lambda_{\text{lat}} = 0.1$	**0.057**	**0.108**	**0.066**	0.127	**76.09**	**75.05**	**74.70**	**72.40**	**86.82**
$\lambda_{\text{lat}} = 0.3$	0.063	0.113	0.072	**0.125**	72.54	72.65	72.96	71.48	85.29
$\lambda_{\text{lat}} = 1.0$	0.060	0.110	0.076	0.129	74.82	73.03	73.40	71.79	85.03

Temporal modeling for detection and forecasting. Table 4 evaluates the temporal trunk and latent world prediction for observed gaze estimation, shift detection, and forecasting. Removing both worsens observed gaze-point accuracy on both datasets and lowers four-class AP/F1. Increasing the latent loss weight also weakens these results relative to the reference, showing that stronger latent supervision does not necessarily improve gaze readout. The largest tested weight still gives higher shift scores than removing the trunk, so the temporal contribution is reduced rather than uniformly lost. The reference weight, $\lambda_{\text{lat}} = 0.1$, gives the lowest observed gaze-point error and highest shift scores among the tested settings and is used in the remaining experiments. The full observed-gaze metrics are reported in Appendix E.2.

Loss terms of video training. Table 5 shows that combining in/out supervision and latent prediction gives the strongest shift detection. Compared with heatmap-only training, latent prediction improves both localization and four-class AP/F1; in/out supervision alone improves shift scores but worsens localization. Combining both improves localization on both datasets. With the trunk retained, adding latent prediction to heatmap and in/out supervision improves both gaze localization and transition detection. This comparison isolates the benefit of latent prediction within the observed clip, supporting its role in learning temporal representations for the detection stage.

Unfreezing the video encoder. Table 6 supports keeping the pretrained video encoder frozen while adapting the temporal and task components. In the comparative run, the encoder remains frozen for the first five epochs, after which its final four transformer blocks and layer normalization are unfrozen and optimized with a reduced learning rate of 1×10^{-6} for the remaining fifteen epochs. Partial unfreezing provides no consistent improvement in gaze estimation and weakens transition classification, with lower four-class AP and macro-F1. Updating the encoder therefore brings no consistent benefit in this comparison. We keep the encoder frozen throughout training, consistent with prior findings in static gaze estimation (Ryan et al., 2025).

Visualization. Figure 4 shows selected examples of observed and future gaze predictions. In the observed sequences, GWM follows target changes in scenes with multiple people and partial

Table 5: Ablation of the loss terms. Shading marks the reference; – indicates that in/out AP is not reported without $\mathcal{L}_{\text{inout}}$.

Loss terms			Gaze-target estimation						Gaze-shift detection	
			VideoAttentionTarget			ChildPlay			Four-class	
$\mathcal{L}_{\text{heat}}$	$\mathcal{L}_{\text{inout}}$	$\mathcal{L}_{\text{lat}}$	AUC↑	L2↓	AP↑	AUC↑	L2↓	AP↑	AP↑	F1↑
✓	✓	✓	**0.973**	**0.057**	**0.948**	**0.974**	**0.066**	0.997	**76.09**	**75.05**
✓	✓	×	0.963	0.068	0.945	0.969	0.073	**0.999**	73.87	73.78
✓	×	×	0.961	0.064	–	0.971	0.070	–	72.71	71.90
✓	×	✓	0.967	0.061	–	0.971	0.069	–	74.55	74.38

Table 6: Effect of unfreezing the encoder. Shading marks the frozen reference.

Variant	Gaze-target estimation				Gaze-shift detection	
	VideoAttentionTarget		ChildPlay		Four-class	
	AUC↑	L2↓	AUC↑	L2↓	AP↑	F1↑
Last 4 blocks unfrozen (50.4M)	**0.974**	0.058	**0.974**	0.068	75.25	69.85
Frozen (reference)	0.973	**0.057**	**0.974**	**0.066**	**76.09**	**75.05**

occlusion. In the forecasting examples, predicted gaze points remain close to the annotated targets as attention shifts, illustrating the model's ability to anticipate changes beyond the observed clip. Additional qualitative results are provided in Appendix F.

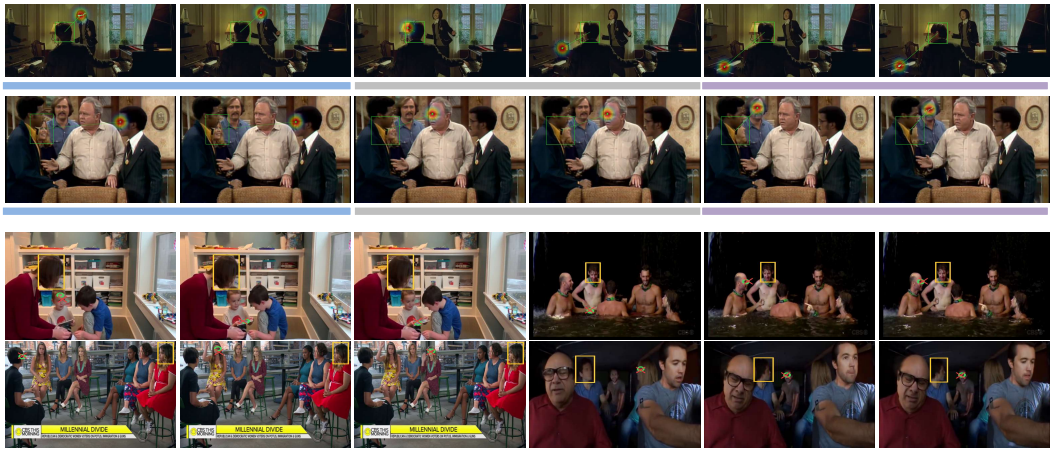

Figure 4: Gaze predictions of GWM. Top: observed heatmaps; bars mark annotated fixation–transition–fixation intervals (blue–gray–purple). Bottom: four observed–future–future triplets. Yellow boxes mark subjects, green circles annotated gaze, and red crosses predictions. Future frames are shown for reference only.

6 Conclusion

We introduced the Gaze World Model (GWM), which models gaze shifts as, in part, a person's response to dynamic changes in the scene. Because this relationship is subtle, GWM co-trains latent scene prediction with person-conditioned gaze and shift supervision on a frozen world-model encoder, sensitizing the model to how scene changes relate to a specific person's attention. The same representation supports gaze estimation and shift detection in the observed clip, and a world predictor extends it to forecast future gaze and shifts. We also augmented VideoAttentionTarget and ChildPlay with event-level gaze-shift annotations through a human-in-the-loop pipeline. Experiments show improved gaze localization, shift detection, and future forecasting over the evaluated baselines, suggesting that modeling how a scene evolves helps explain where a person looks and when their attention shifts. Future work can extend this approach to longer forecast horizons, richer target identities, and a broader range of interaction settings.

AI USE STATEMENT

We used Gemini (specifically Gemini 3.5 Flash) to assist with the event-level semantic verification stage of our gaze-shift data curation pipeline (Section 4 and Appendix C). Gemini serves as a provides verification judgment and does not create the source gaze coordinates. Appendix C describes the automated checks and human in-loop audits. We also used generative AI tools to assist coding, editing manuscript text, refining formatting, checking mathematical formulations against code implementations, searching literature, and drafting figure designs. The authors manually reviewed the AI-assisted work and are responsible for the final text, mathematical statements, experimental claims, and artifacts .

ETHICS STATEMENT

This research uses VideoAttentionTarget (VAT) and ChildPlay for video tasks, together with GazeFollow and MVGT for the image stage. Access and use of these benchmarks comply with their respective licenses and terms of service. ChildPlay involves recordings of pediatric subjects in play environments; our study performs non-invasive computational modeling only. We use previously released research datasets under their applicable access and usage terms.

Our semi-automated data curation pipeline incorporates Gemini 3.5 Flash for event-level candidate filtering. The only data sent to the external verifier were contact sheets rendered from these public datasets, which may show identifiable faces; no other personal information was sent. Instances flagged by automated filters or ambiguity thresholds were handled through explicit abstention or human audit (Section 4 and Appendix C).

Social impact and potential dual-use: Modeling human gaze shifts and attention dynamics has beneficial applications in human-robot collaboration, cognitive science, developmental psychology, and accessible assistive interfaces. However, subject-conditioned gaze-following models also present surveillance and privacy risks if deployed for unauthorized tracking, behavioral profiling, or covert human monitoring across camera networks. Furthermore, estimated gaze targets reflect visual orientation and do not directly entail internal cognitive intent, emotional state, or mental health status. We caution against applying these models to unconsented surveillance or high-stakes decision-making without explicit user consent, strict privacy safeguards, and human oversight.

REPRODUCIBILITY STATEMENT

We describe the architecture and training in Sections 3 and 5.1, and the data construction in Appendix C. Upon publication, we will release the training and inference code, configurations and initialization settings used for the reported models, and the trained model weights. The data release will include the rule configurations, verifier prompts, annotation-version identifiers, and event-to-window construction settings needed to reproduce the reported splits, subject to the source datasets' distribution terms.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024.

Pieter Blignaut. Fixation identification: The optimum threshold for a dispersion algorithm. *Attention, Perception, & Psychophysics*, 71(4):881–895, 2009. doi: 10.3758/APP.71.4.881.

Xu Cao, Houze Yang, Vipin Gunda, Zhongyi Zhou, Tianyu Xu, Adarsh Kowdle, Inki Kim, and James M. Rehg. Gaze target estimation anywhere with concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31304–31315, 2026.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations*, 2026.

Lu Chen, Yizhou Wang, Shixiang Tang, Qianhong Ma, Tong He, Wanli Ouyang, Xiaowei Zhou, Hujun Bao, and Sida Peng. Egoagent: A joint predictive agent model in egocentric worlds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6970–6980, 2025.

Eunji Chong, Nataniel Ruiz, Yongxin Wang, Yun Zhang, Agata Rozga, and James M. Rehg. Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 397–412, 2018.

Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M. Rehg. Detecting attended visual targets in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5396–5406, 2020. doi: 10.1109/CVPR42600.2020.00544.

Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, Fengli Xu, and Yong Li. Understanding world or predicting future? a comprehensive survey of world models. *ACM Computing Surveys*, 58(3): 1–38, 2025. doi: 10.1145/3746449.

Alexandra Frischen, Andrew P. Bayliss, and Steven P. Tipper. Gaze cueing of attention: Visual attention, social cognition, and individual differences. *Psychological Bulletin*, 133(4):694–724, 2007. doi: 10.1037/0033-2909.133.4.694.

Antonino Furnari and Giovanni Maria Farinella. Rolling-unrolling LSTMs for action anticipation from first-person video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43 (11):4021–4036, 2021. doi: 10.1109/TPAMI.2020.2992889.

Gemini Team. Gemini 3.8 Flash model card. https://deepmind.google/models/model-cards/gemini-3-8-flash/, 2026.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Mehrdad Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Anshul Gupta, Samy Tafasca, Arya Farkhondeh, Pierre Vuillecard, and Jean-Marc Odobez. MTGS: A novel framework for multi-person temporal gaze following and social gaze prediction. In *Advances in Neural Information Processing Systems*, pp. 15646–15673, 2024. doi: 10.52202/079017-0500.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025. doi: 10.1038/s41586-025-08744-2.

Mary Hayhoe and Dana Ballard. Eye movements in natural behavior. *Trends in Cognitive Sciences*, 9(4):188–194, 2005. doi: 10.1016/j.tics.2005.02.009.

Roy S. Hessels, Antje Nuthmann, Marcus Nyström, Richard Andersson, Diederick C. Niehorster, and Ignace T. C. Hooge. The fundamentals of eye tracking part 1: The link between theory and research question. *Behavior Research Methods*, 57(1):16, 2025. doi: 10.3758/s13428-024-02544-8.

Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer. *Eye Tracking: A Comprehensive Guide to Methods and Measures*. Oxford University Press, Oxford, UK, 2011. ISBN 9780199697083.

Ignace T. C. Hooge, Antje Nuthmann, Marcus Nyström, Diederick C. Niehorster, Gijs A. Holleman, Richard Andersson, and Roy S. Hessels. The fundamentals of eye tracking part 2: From research question to operationalization. *Behavior Research Methods*, 57(2):73, 2025. doi: 10.3758/s13428-024-02590-2.

Ignace T. C. Hooge, Marcus Nyström, Diederick C. Niehorster, Richard Andersson, Tom Foulsham, Antje Nuthmann, and Roy S. Hessels. The fundamentals of eye tracking part 6: Working with areas of interest. *Behavior Research Methods*, 58(3):65, 2026. doi: 10.3758/s13428-025-02937-3.

Chien-Ming Huang, Sean Andrist, Allison Sauppé, and Bilge Mutlu. Using gaze patterns to predict task intent in collaboration. *Frontiers in Psychology*, 6:1049, 2015. doi: 10.3389/fpsyg.2015.01049.

Chenyi Kuang and Nakul Agarwal. SAGE: Synchronized action-gaze recognition and anticipation for human behavior understanding. In *European Conference on Computer Vision*, 2026.

Bolin Lai, Miao Liu, Fiona Ryan, and James M. Rehg. In the eye of transformer: Global-local correlation for egocentric gaze estimation. In *British Machine Vision Conference*, 2022.

Michael F. Land. Eye movements and the control of actions in everyday life. *Progress in Retinal and Eye Research*, 25(3):296–324, 2006. doi: 10.1016/j.preteyeres.2006.01.002.

Jia Li, Wenjie Zhao, Shijian Deng, Bolin Lai, Yuheng Wu, Ruijia Chen, Jon E. Froehlich, Yuhang Zhao, and Yapeng Tian. ARGaze: Autoregressive transformers for online egocentric gaze estimation. *arXiv preprint arXiv:2602.05132*, 2026.

Yin Li, Miao Liu, and James M. Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European Conference on Computer Vision*, pp. 639–655, 2018. doi: 10.1007/978-3-030-01228-1_38.

Dongze Lian, Zehao Yu, and Shenghua Gao. Believe it or not, we know what you are looking at! In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pp. 35–50, 2018.

Russell Mendonca, Shikhar Bahl, and Deepak Pathak. Structured world models from human videos. In *Robotics: Science and Systems*, 2023.

Qiaomu Miao, Vivek Raju Golani, Jingyi Xu, Progga Paromita Dutta, Minh Hoai, and Dimitris Samaras. Multi-view gaze target estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5371–5381, 2025. doi: 10.1109/ICCV51701.2025.00510.

Qiaomu Miao, Haoyu Wu, Jingyi Xu, Minh Hoai, and Dimitris Samaras. OmniGF: A dual-branch vision-language framework for unified gaze following. *arXiv preprint arXiv:2605.26399*, 2026.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mahmoud Assran, Koustuv Sinha, Michael Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-JEPA 2.1: Unlocking dense features in video self-supervised learning. *arXiv preprint arXiv:2603.14482*, 2026.

Diederick C. Niehorster, Marcus Nyström, Roy S. Hessels, Jeroen S. Benjamins, Richard Andersson, and Ignace T. C. Hooge. The fundamentals of eye tracking, part 7: Determining data quality. *Behavior Research Methods*, 58(7):183, 2026. doi: 10.3758/s13428-026-03039-4.

Qwen Team. Qwen3.8-27B. `https://huggingface.co/Qwen/Qwen3.8-27B`, 2026.

Adrià Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Fiona Ryan, Ajay Bati, Sangmin Lee, Daniel Bolya, Judy Hoffman, and James M. Rehg. Gazelle: Gaze target estimation via large-scale learned encoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28874–28884, 2025. doi: 10.1109/CVPR52734.2025.02689.

Dario D. Salvucci and Joseph H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications (ETRA)*, pp. 71–78, 2000.

Junhui She, Fei Wang, Kun Li, Yiqi Nie, Yuxin Liu, Zhangling Duan, and Xun Yang. Textgaze: Prompting gaze target estimation with textual scene cues. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2026.

Qiuhong Shen, Shihua Zhang, Yue Liao, Qi Li, Zhenxiong Tan, Shizun Wang, Shuicheng Yan, and Xinchao Wang. World action models: A survey. *arXiv preprint arXiv:2606.20781*, 2026.

Samy Tafasca, Anshul Gupta, and Jean-Marc Odobez. Childplay: A new benchmark for understanding children's gaze behaviour. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20935–20946, 2023. doi: 10.1109/ICCV51070.2023.01914.

Samy Tafasca, Anshul Gupta, and Jean-Marc Odobez. Sharingan: A transformer architecture for multi-person gaze following. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2008–2017, 2024. doi: 10.1109/CVPR52733.2024.00196.

Michael Tomasello and Michael Jeffrey Farrar. Joint attention and early language. *Child development*, pp. 1454–1463, 1986.

Mengmeng Wang, Zeyi Huang, Xiangjie Kong, Guojiang Shen, Guang Dai, Jingdong Wang, and Yong Liu. Action detail matters: Refining video recognition with local action queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19132–19142, 2025.

Zhoutong Ye, Chengwen Zhang, Zhaibin Cui, Mingze Sun, Jiaqi Liu, Xiangwu Li, Qingyang Wan, Chang Liu, Xutong Wang, Huan-ang Gao, Yu Mei, Chun Yu, and Yuanchun Shi. PaGE: Towards practical human-level gaze target estimation. *arXiv preprint arXiv:2607.04860*, 2026.

Mengmi Zhang, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Jiashi Feng. Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4372–4381, 2017. doi: 10.1109/CVPR.2017.377.

APPENDIX

A Extended Related Work **16**

A.1 Gaze-event definitions and eye-tracking methodology 16

A.2 Gaze and action anticipation 16

A.3 Video gaze architectures and subject conditioning 16

B Model Details **16**

C Gaze-Shift Data Engine **17**

D Evaluation Metrics **18**

D.1 Temporal metrics 18

E Supplementary Experiments **19**

E.1 Image-stage spatial validation 19

E.2 Full metrics for the temporal ablation 19

E.3 Class weighting of the shift detection loss 19

E.4 Temporal sampling and learning rate 20

F Additional Qualitative Examples **21**

G Limitations and Future Work **23**

A Extended Related Work

A.1 Gaze-event definitions and eye-tracking methodology

A gaze event is defined operationally rather than read directly from a raw signal. Established eye-tracking methodology separates continuous ocular kinematics from discrete behavioral events (Salvucci & Goldberg, 2000; Holmqvist et al., 2011; Hessels et al., 2025). Classical velocity- and dispersion-threshold methods isolate fixations and saccades from point-of-regard streams (Salvucci & Goldberg, 2000; Blignaut, 2009). In dynamic scenes, however, target transitions also depend on moving object boundaries, scene geometry, and target visibility (Hayhoe & Ballard, 2005; Niehorster et al., 2026), motivating criteria such as minimum dwell times, transition-duration bounds, and measurement noise floors (Hooge et al., 2025; 2026). These protocols assume high-frequency eye-tracker signals. The third-person datasets used here do not provide synchronized high-frequency eye-tracking signals, so target-level events must instead be defined through the temporal persistence of target identity, supported by instance tracking and semantic verification. Out-of-frame gaze must likewise be treated as an explicit visibility state rather than missing data, which does not by itself distinguish one off-screen target from another (Chong et al., 2020; Tafasca et al., 2023).

A.2 Gaze and action anticipation

Visual anticipation has generally been formulated as predicting a future gaze coordinate or an upcoming action label. In egocentric vision, early work forecast prospective fixation points from wearable-camera streams (Zhang et al., 2017); later methods connected gaze to action recognition and anticipation (Li et al., 2018; Lai et al., 2022) or modeled gaze trajectories autoregressively (Li et al., 2026). SAGE jointly models human–object interactions and visual attention across egocentric and exocentric views (Kuang & Agarwal, 2026), while large-scale video benchmarks define anticipation through symbolic action categories (Grauman et al., 2022; Furnari & Farinella, 2021). We complement gaze-coordinate prediction with explicit person-involving transition labels for a specified subject. Detection classifies these transitions in observed video, while forecasting predicts whether one will occur in the future.

A.3 Video gaze architectures and subject conditioning

Video gaze models have largely extended frame-wise gaze estimation by aggregating independently encoded frame features. Recurrent temporal processing supports out-of-frame gaze prediction (Chong et al., 2020), while later systems incorporate multi-agent reasoning (Tafasca et al., 2024), temporal-social classification (Gupta et al., 2024), multi-view fusion (Miao et al., 2025), and frozen foundation encoders (Ryan et al., 2025; Ye et al., 2026). Subject specification has also expanded beyond head bounding boxes to textual and multimodal prompts (Cao et al., 2026; She et al., 2026; Miao et al., 2026). Nevertheless, applying spatial estimators frame by frame can produce temporal jitter (Chong et al., 2020), and post-hoc smoothing can suppress genuine rapid transitions (Li et al., 2026). This motivates learning temporal representations before decoding gaze rather than treating temporal context solely as an aggregation or filtering step.

B Model Details

Head-crop encoding. For each observed frame, the head crop of the specified person is cut from the frame using its head box and resized to 256×256. The head-crop clip passes through the same frozen V-JEPA 2.1 encoder as the scene clip in a separate forward pass, so the two streams share encoder weights and no additional head encoder is trained. With the same two-frame tubelets, each head tubelet yields $P_h = 16 \times 16 = 256$ patch features of width C. SIM then projects these features to width D and places the head tokens inside the head-box rectangle through their RoPE coordinates (Section 3.2).

Baseline adaptation. Gaze-LLE and PaGE are image models without a shift-detection output. We keep each model frozen and extract, per frame, its person-conditioned scene features (after SIM for PaGE; the input to the heatmap head for Gaze-LLE). These features are pooled with the model's own

Table 7: Terminology for annotated events, benchmark labels, and model outputs.

Term	Meaning in this work
Semantic target	The person, object, or region to which gaze is assigned; object keys do not guarantee persistent instance identity.
Person identity	A person identifier propagated during annotation, distinguishing different people from body parts of the same person.
`other`	Objects, background regions, and out-of-frame gaze; it does not identify a particular off-screen target.
Event onset / offset	The annotated start and end of the transition interval between stable targets A and B.
Window label	A gaze-shift event is assigned to a window if the transition starts within that window, even if it ends after the window. Observed windows have four labels; future windows have binary labels.
No in-scope shift	A valid window contains none of the three person-involving transition types.
Unverified	Event status or required annotations remain unresolved after review; affected windows are excluded, not labeled negative.

predicted heatmap and concatenated with its predicted gaze point, in/out probability, and the head box; no ground-truth gaze enters this input. A trainable temporal classifier (two Transformer layers with sinusoidal time encoding, masked mean pooling, and an MLP) maps the per-frame vectors to the four window classes. SAGE has no public code, so we re-implement its gaze-anticipation branch without the action branch, which requires action labels that VAT and ChildPlay lack. Its human–object context tokens are replaced by pooled scene tokens of the gaze detector, and one anticipation step corresponds to one two-frame tubelet. The gaze detector is either Sharingan, initialized from its released VAT checkpoint, or PaGE ViT-B, and is applied frame by frame. For SAGE with Sharingan, the detector is frozen and only the anticipation module is trained. For SAGE with PaGE, the detector is fine-tuned jointly with the anticipation module following SAGE's two-step schedule: 5 epochs on current-gaze labels, then 20 epochs with future-gaze labels added. SAGE has no shift output, so for binary shift forecasting we freeze the trained SAGE model and train a shift head on its per-step anticipation tokens, with the same training data, loss, and optimization as the forecasting head of GWM.

C Gaze-Shift Data Engine

Table 7 defines the terms used in Section 3.1. We then detail the data engine summarized in Section 4 and Figure 5.

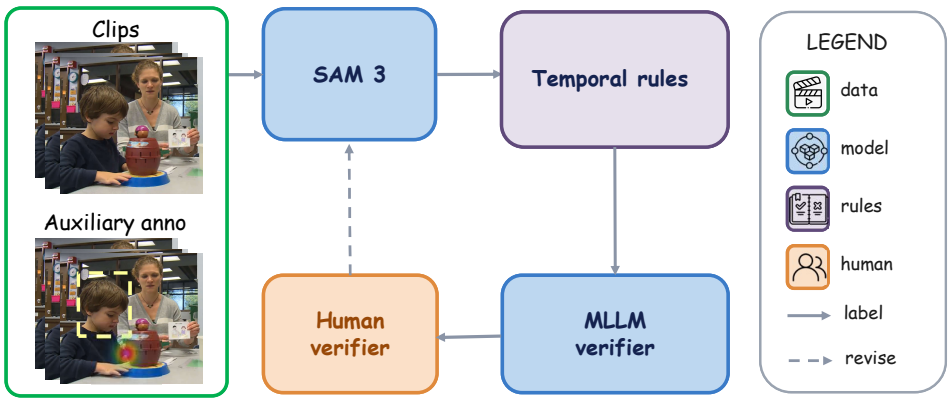

Figure 5: The three-stage gaze-shift data engine.

Stage 1: rule-based candidate generation. For each annotated person and frame, the rule stage assigns a gaze target. Out-of-frame status is taken from the source annotation, the coordinate $(-1, -1)$ in VAT and the `outside_frame` state in ChildPlay, and is never inferred from a missing

coordinate. For in-frame gaze, SAM 3 (Carion et al., 2026) provides human-instance masks from a "person" text prompt and a target mask from the annotated gaze point used as a point prompt. A target that overlaps a person mask becomes a human target with that person's identifier, and any other target becomes an object target. The resulting sequence of target keys is run-length encoded, and runs of the same target separated by a short gap are merged. Two stable runs that each last at least T_{dwell} and are separated by a transition of at most T_{trans} form a candidate shift (Table 8). The rule stage abstains when mask agreement is low, when more than half of the frames have low confidence, or when more than 30% of the frames lack a mask.

Stage 2: multimodal verification. Each candidate is rendered as one JPEG contact sheet whose tiles cover the frames before, during, and after the transition and overlay the head box, the annotated gaze point, the SAM 3 target contour, and the target key. Gemini 3.5 Flash judges the candidate at the event level and returns a schema-constrained output with a shift decision, a confidence, and normalized pre-shift and post-shift targets. Three constraints keep the verifier from overriding the geometric evidence: person identifiers must be copied from the rule stage, and the verifier abstains when no unique identifier is available; in-frame and out-of-frame status follows the source annotation; and a held or worn item remains an object target rather than being relabeled as its holder. A positive decision confirms the candidate, a negative decision retains it as a hard negative, and an error or abstention leaves it unverified pending human review.

Stage 3: human verification. Candidates are routed to human reviewers when mask agreement is low, the gaze displacement is small, a temporal measurement lies close to a rule threshold, the semantic target is ambiguous, or the verifier abstains or conflicts with the rule evidence. Reviewers inspect the original frames and the contact sheet, and confirm or correct the shift decision, the two endpoint targets, and the transition boundaries. Cases that remain unresolved are marked as unverified and excluded from training and evaluation. Each audit randomly samples 15% of all annotations, requires a sampled error rate below 1%, and is followed by manual correction of the errors found.

D Evaluation Metrics

Table 9 lists the metric families and their evaluation settings. Spatial accuracy follows the official protocol of each benchmark. Four-class gaze-shift AP is micro-averaged across all four classes. Tables 1 and 3 report detection and forecasting results; this appendix also defines the temporal metrics.

D.1 Temporal metrics

Frame-level L2 and AUC score each frame independently, so they cannot tell whether a prediction is stable during a fixation or whether it moves when the gaze moves. The metrics in Table 2 measure these two properties on the prediction sequence of each test track at 8 FPS. Let $\hat{g}_t$ be the arg-max of the predicted heatmap at frame t and g_t^* the annotated gaze point, both in normalized image coordinates. All models provide predictions on the same evaluated frames; a frame is valid when g_t^* lies inside the image. Segmentation, thresholds, and aggregation are identical for every model.

Segments. Segments are derived from the annotated gaze points only, so every model is evaluated on the same frames. A fixation segment is a maximal run of at least four consecutive valid frames in which g_t^* moves by less than $\tau = 0.02$ between adjacent frames. A movement segment is a maximal run of at least two consecutive valid frames in which g_t^* moves by at least τ between adjacent frames.

Table 8: Main rule-filter thresholds.

Parameter	Value	Function
Stable fixation (T_{dwell})	133 / 150 ms	fixation per endpoint (VAT / ChildPlay)
Maximum transition (T_{trans})	1,000 ms	maximum interval between stable endpoints
Same-target merge gap	67 ms	maximum gap for merging runs of one target

Table 9: Metric families and their evaluation scope.

Metric family	Image	Observed video	Future video
Spatial accuracy	✓	✓	✓
Temporal dynamics		✓	
Gaze-shift events		✓	✓

Movement segments therefore include any motion of the annotated point, not only the verified gaze shifts of Appendix C.

Jitter and Flicker. Before computing Jitter and Flicker, we aggregate baseline predictions over the same two-frame intervals to match GWM's temporal output resolution. Let $\mathcal{S}$ be the set of adjacent frame pairs $(t-1, t)$ within fixation segments, pooled over all tracks. Then

$$\text{Jitter} = \frac{1}{|\mathcal{S}|} \sum_{(t-1,t)\in\mathcal{S}} \|\hat{g}_t - \hat{g}_{t-1}\|_2, \qquad \text{Flicker} = \frac{1}{|\mathcal{S}|} \sum_{(t-1,t)\in\mathcal{S}} \mathbb{I}\left(\|\hat{g}_t - \hat{g}_{t-1}\|_2 > \theta\right), \quad (2)$$

with $\theta = 0.10$. Jitter measures the average frame-to-frame change, and Flicker the fraction of large jumps.

Aligned residual. For a movement segment $[s, e]$ in a track of n frames, we compare the annotated trajectory with the prediction shifted by k frames and keep the best offset within two frames (250 ms):

$$E_{[s,e]}(k) = \frac{1}{|V_k|} \sum_{t\in V_k} \|\hat{g}_{t+k} - g^*_t\|_2, \qquad r_{[s,e]} = \min_{k\in\{-2,\dots,2\}} E_{[s,e]}(k), \quad (3)$$

where $V_k = \{t \in [s, e] : 0 \le t + k < n\}$. The reported r is the mean of $r_{[s,e]}$ over all movement segments. The offset search removes short delays, so r measures whether the prediction follows the path of the gaze rather than its latency. A low Jitter with a low r indicates a prediction that is stable during fixations and still follows the gaze when it moves; a low Jitter with a high r indicates that stability was obtained by suppressing movement.

E Supplementary Experiments

This section collects the expanded comparisons and diagnostic experiments that support the main video results.

E.1 Image-stage spatial validation

Table 10 reports image-level GTE on the official held-out test annotations of VAT and ChildPlay with ground-truth head boxes. The comparison provides a spatial reference for the later video stages.

E.2 Full metrics for the temporal ablation

Table 11 reports the full observed-gaze and shift-detection metrics for the variants in Table 4.

E.3 Class weighting of the shift detection loss

We compare four class-weighting strategies for the shift detection loss, keeping the video-stage initialization and all other settings fixed. Table 12 reports results separately for each dataset. Balanced weighting leads every reported metric on both datasets. Uniform weighting ranks second on VideoAttentionTarget in accuracy, F1, and AP, while inverse-frequency weighting ranks second on ChildPlay in accuracy, F1, and M-F1 but has the lowest AP. Thus, balanced weighting gives the most consistent performance across datasets and metrics in this comparison.

Table 10: Image-stage gaze-target estimation on VAT and ChildPlay. AUC is not reported (–) for the VLMs, which output gaze points without a probability heatmap.

Model	VideoAttentionTarget			ChildPlay		
	AUC ↑	L2 ↓	AP ↑	AUC ↑	L2 ↓	AP ↑
Specialized gaze models						
Gaze-LLE ViT-B	0.946	0.087	0.883	0.956	0.110	0.997
Gaze-LLE ViT-L	0.952	0.084	0.886	0.959	0.101	0.995
PaGE ViT-B	0.971	0.069	0.947	0.955	0.086	0.997
PaGE ViT-H+	**0.973**	0.066	0.940	0.958	0.082	0.996
Vision-language and generalist models						
Qwen3.8 27B	–	0.159	0.784	–	0.147	0.979
Gemini 3.8 Flash	–	0.146	0.824	–	0.119	0.981
Ours						
GWM (ViT-B)	0.962	0.069	0.943	0.963	0.074	0.995
GWM (ViT-L)	0.969	0.066	0.950	0.967	0.072	0.998
GWM (ViT-g)	**0.973**	**0.063**	**0.954**	**0.971**	**0.069**	**0.999**

Table 11: Full observed-gaze and shift-detection metrics for the temporal modeling ablations in Table 4.

Variant	Gaze-target estimation						Gaze-shift detection	
	VideoAttentionTarget			ChildPlay			Four-class (%)	
	AUC↑	L2↓	AP↑	AUC↑	L2↓	AP↑	AP↑	F1↑
−temporal trunk	0.966	0.069	0.937	0.967	0.076	**0.999**	71.25	69.97
−world predictor	0.970	0.068	0.939	0.971	0.073	0.998	73.92	73.71
$\lambda_{\text{lat}} = 0.1$	**0.973**	**0.057**	**0.948**	**0.974**	**0.066**	0.997	**76.09**	**75.05**
$\lambda_{\text{lat}} = 0.3$	0.965	0.063	0.931	0.967	0.072	**0.999**	72.54	72.65
$\lambda_{\text{lat}} = 1.0$	0.970	0.060	0.945	0.964	0.076	**0.999**	74.82	73.03

Table 12: Class weighting of the four-class shift detection loss on each dataset. Acc. and M-F1 are four-class accuracy and macro-F1; F1 and AP are binary shift scores. All scores are percentages.

Weighting	VideoAttentionTarget				ChildPlay			
	Acc.↑	F1↑	M-F1↑	AP↑	Acc.↑	F1↑	M-F1↑	AP↑
Balanced	**79.80**	**79.20**	**72.95**	**85.33**	**75.82**	**82.75**	**77.26**	**87.99**
Uniform	78.94	77.85	69.80	85.12	71.52	76.67	73.14	87.53
Inverse-frequency	73.89	75.27	67.76	84.85	74.90	80.55	73.60	84.65
Focal ($\gamma = 2$)	76.80	76.59	70.68	84.08	73.39	75.41	72.93	86.00

E.4 Temporal sampling and learning rate

Table 13 reports temporal-sampling settings and a learning-rate variant under the official dense video test protocol. The reference (16 frames at 8 FPS) has the lowest L2 and highest AUC on both datasets. Among the other sampling settings, 32 frames at 8 FPS gives the lowest L2 (tied on VideoAttentionTarget with 16 frames at 16 FPS); frame-rate effects vary across datasets. Halving the learning rate gives worse scores than the reference on both datasets. The sampling settings use different valid-window sets, so the comparisons are not strictly paired.

Table 13: Temporal sampling and learning-rate sensitivity for gaze-target estimation.

	VideoAttentionTarget		**ChildPlay**	
Setting	L2↓	AUC↑	L2↓	AUC↑
8f @ 4 FPS	0.073	0.962	0.078	0.973
8f @ 8 FPS	0.068	0.965	0.075	0.972
16f @ 4 FPS	0.066	0.964	0.075	0.963
16f @ 16 FPS	0.065	0.970	0.074	0.967
32f @ 8 FPS	0.065	0.966	0.070	0.968
32f @ 16 FPS	0.069	0.970	0.073	0.965
Main (16f @ 8 FPS)	**0.057**	**0.973**	**0.066**	**0.974**
Learning rate $\times 0.5$	0.064	0.969	0.075	0.968

F Additional Qualitative Examples

Figures 6–8 compare GWM with external baselines on observed gaze-target estimation, gaze-shift detection, and future gaze-target forecasting. The localization examples show predictions across consecutive frames in scenes with multiple people and competing targets. The detection examples show other-to-human and human-to-human transitions; baseline errors include confusing transition types and predicting no shift. The forecasting examples compare GWM with SAGE with Sharingan and SAGE with PaGE on the same clips, showing how predicted gaze locations track changes in the annotated target beyond the observation window. These selected cases illustrate model behavior in specific scenes and complement the quantitative evaluation.

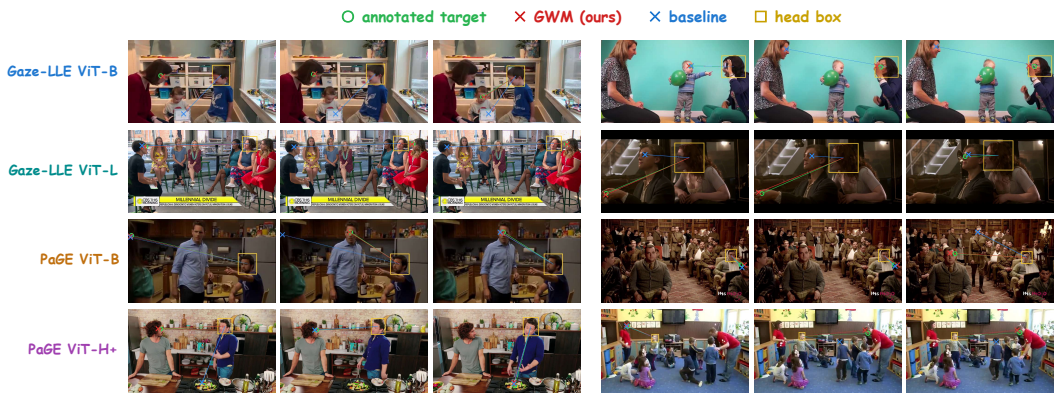

Figure 6: Gaze-target estimation on consecutive observed frames. Each row compares GWM (red crosses) with one baseline (blue crosses); green circles mark annotated targets and yellow boxes mark the specified subjects.

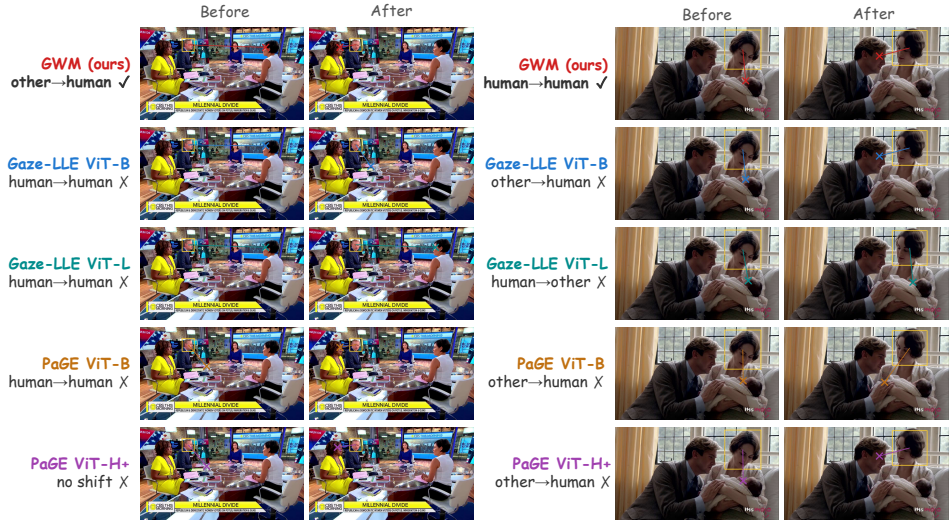

Figure 7: Gaze-shift detection in two observed clips. Each row shows the frames before and after the annotated transition, together with the transition type predicted by each model (✓ correct, × incorrect).

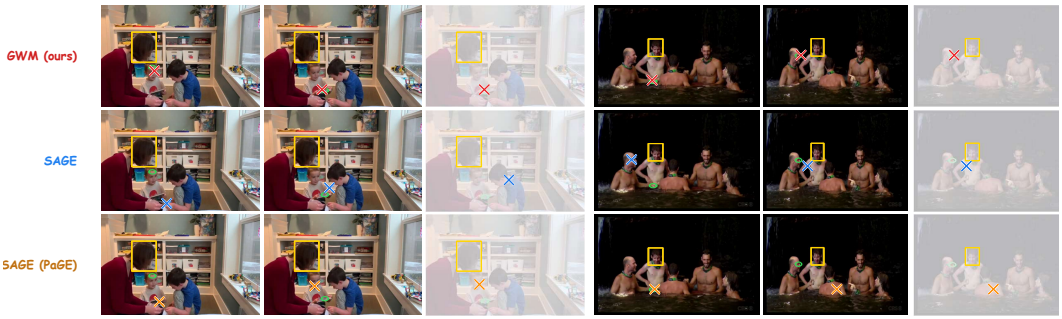

Figure 8: Gaze-target forecasting compared with SAGE with Sharingan and SAGE with PaGE. Each three-column example shows an observed frame, a future frame at the shift onset, and a later future frame (faded). Green circles mark annotated gaze targets; crosses show predictions from GWM (red), SAGE with Sharingan (blue), and SAGE with PaGE (orange). Yellow boxes mark the specified subjects. Future frames are shown for reference only and are not model inputs.

G LIMITATIONS AND FUTURE WORK

Temporal scope. The present experiments use finite observation windows and fixed forecast horizons, with gaze outputs produced at two-frame tubelet resolution. Future work could extend the model to longer context, multiple forecast horizons, and finer temporal readout, and evaluate event-boundary localization alongside window-level detection and forecasting.

Subject conditioning and target representation. An important next step is to evaluate subject conditioning with automatically detected and temporally tracked head boxes, including sensitivity to crop quality. Adding explicit target representations and temporal object association could support instance-level gaze transitions, distinguish objects of the same category, and connect latent future predictions to interpretable target identities. Richer treatment of off-screen targets would further broaden the event vocabulary.

Annotation and evaluation. Further work could expand the human audit across datasets and interaction settings, quantify agreement between annotators and the semantic verifier, and evaluate additional transition types under a common protocol. Such extensions would complement the current person-centered detection taxonomy and binary forecasting task.