VANTAGE: TEMPORALLY STRUCTURED PREDICTIVE UNCERTAINTY FOR PROBABILISTIC EGOCENTRIC GAZE ANTICIPATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Egocentric gaze anticipation is inherently uncertain, yet existing methods predominantly predict a single deterministic future sequence, leaving plausible future variations unmodeled, the structure of predictive uncertainty uncharacterized, and its reliability information unused. Our key insights are that (i) the conditional variability of future gaze is temporally structured: the uncertainty varies along the anticipation horizon, and (ii) predictive uncertainty manifests in trajectory-level and gaze heatmap-level forms. We therefore reformulate the task as conditional probabilistic forecasting and propose **VANTAGE**, a temporally structured variational framework that unifies latent modeling and two-level uncertainty estimation over a shared interval-based horizon decomposition, with the estimated uncertainty directly guiding temporal attention during decoding. Experiments on three datasets demonstrate the effectiveness of the proposed temporally structured modeling and uncertainty estimation, while revealing different roles of the two forms. These results support predictive uncertainty as both a measure of prediction reliability and a signal for temporal reasoning.

1 INTRODUCTION

Egocentric gaze anticipation aims to predict the wearer's future gaze sequence from historical observations, providing an informative early signal for wearable assistance, AR/VR, human–robot interaction, and embodied systems (Shafti et al., 2019; Belardinelli, 2024; Li et al., 2026). Prior works have improved gaze anticipation through richer feature representations, including spatio-temporal features, multimodal cues, generated future frames, and 3D scene representations (Zhang et al., 2018; Lai et al., 2024b; Yun et al., 2025; Ryan et al., 2026). Despite these advances, they predominantly produce and evaluate a single future gaze sequence for each observation.

However, natural gaze control depends on uncertain world states and evolving behavioral goals, and substantial variability persists even under the same dynamic stimulus (Goettker et al., 2025; Hayhoe & Matthis, 2018; Dorr et al., 2010); Prior work has also noted the existence of multiple plausible future gaze trajectories (Ryan et al., 2026). Egocentric gaze anticipation is therefore inherently uncertain: historical observations constrain but do not uniquely determine future gaze. Yet standard datasets (Grauman et al., 2022; 2024) provide only one realized future gaze sequence for each training example, and such supervision does not explicitly characterize the distribution of plausible futures. This motivates a probabilistic formulation of egocentric gaze anticipation.

Our key insight is twofold. First, the conditional variability of future gaze is temporally structured: it varies along the anticipation horizon, with different intervals exhibiting different degrees of uncertainty. Second, this predictive uncertainty manifests in two complementary forms: variability across plausible gaze trajectories and spatial dispersion within each individual gaze heatmap. However, latent distribution parameters alone do not directly quantify this uncertainty in output space, as latent variability is transformed nonlinearly by the decoder (Capellera et al., 2025). Taken together, these insights motivate modeling future gaze variability with temporal structure and characterizing its trajectory- and heatmap-level output-space uncertainty at the same interval granularity.

We propose **VANTAGE**, a temporally structured variational framework that unifies latent modeling, uncertainty estimation, and temporal attention over a shared interval-based horizon decomposition.

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.

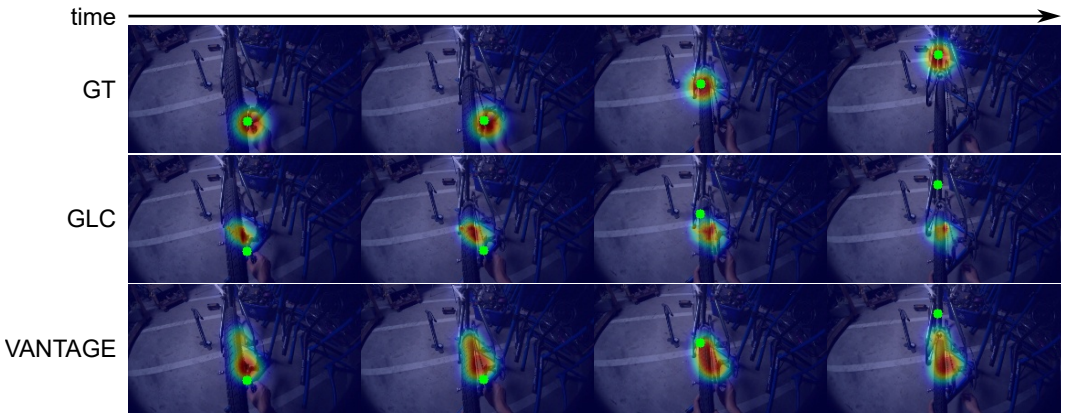


Figure 1: Qualitative comparison: GLC (Lai et al., 2024a) and VANTAGE on EgoExo-4D.

Each interval is assigned a distinct latent representation and corresponding output-space uncertainty estimates at both trajectory- and heatmap-level. Crucially, these interval-specific uncertainty estimates are obtained by estimator trained with MC supervision, and passed to modulate temporal attention, enabling weighted uncertainty-aware aggregation according to their reliability. At inference, they are computed in a single forward pass without additional stochastic decoding.

Our contributions are threefold: **First**, we reformulate egocentric gaze anticipation as conditional probabilistic forecasting over future gaze heatmap sequences and provide an evaluation protocol for both distributional prediction and uncertainty reliability. **Second**, we introduce temporally structured latent representations and interval-wise output-space uncertainty at trajectory and heatmap levels over a shared horizon decomposition, to model future gaze variability over time and quantify it directly in output space. **Third**, we make predictive uncertainty available before decoding, giving it a dual role in characterizing reliability and guiding temporal reasoning for uncertainty-aware prediction.

2 Related Work

Egocentric Gaze Modeling. Egocentric gaze modeling encompasses two related but distinct tasks: egocentric gaze estimation which predicts gaze in observed frames, and egocentric gaze anticipation which predicts gaze in unseen future frames. Gaze estimation has been studied using CNNs, recurrent models, and Transformers that capture spatio-temporal, human–object, and multimodal context (Li et al., 2013; Huang et al., 2018; 2020; Li et al., 2021; Lai et al., 2022; 2024a; Li et al., 2023). Gaze anticipation instead predicts future gaze from historical observations. Prior approaches have incorporated future-frame synthesis (Zhang et al., 2018), audio-visual cues (Lai et al., 2024b), ego–exo correspondence (Qiu et al., 2026), and 3D scene representations (Yun et al., 2025; Ryan et al., 2026). These methods predominantly produce and evaluate a single future gaze sequence in coordinate or heatmap form. Other related generative scanpath models have explored stochastic gaze prediction to capture variability (Wang et al., 2024; Cartella et al., 2025), but focus on scanpath generation rather than probabilistic gaze anticipation. Our work instead models a conditional distribution over future gaze sequences together with temporally structured predictive uncertainty.

Probabilistic Trajectory Modeling. Probabilistic trajectory forecasting models multiple plausible futures using GANs (Gupta et al., 2018; Kosaraju et al., 2019; Dendorfer et al., 2021), conditional VAEs (Salzmann et al., 2020; Yuan et al., 2021; Lee et al., 2022), and diffusion models (Zhu et al., 2023; Bae et al., 2024; Yang et al., 2024). These methods inspire our probabilistic formulation, but address a different prediction objective: trajectory forecasting predicts future physical states, whereas egocentric gaze anticipation predicts the transition of task-conditioned visual attention in the context of a rapidly changing egocentric perspective.

Uncertainty-Aware Modeling. Predictive uncertainty is commonly estimated through stochastic predictions, ensembles, or explicit uncertainty models, and can support reliability assessment and selective prediction (Li et al., 2024; Liu et al., 2024; Bao et al., 2023; Wang et al., 2025; Marx

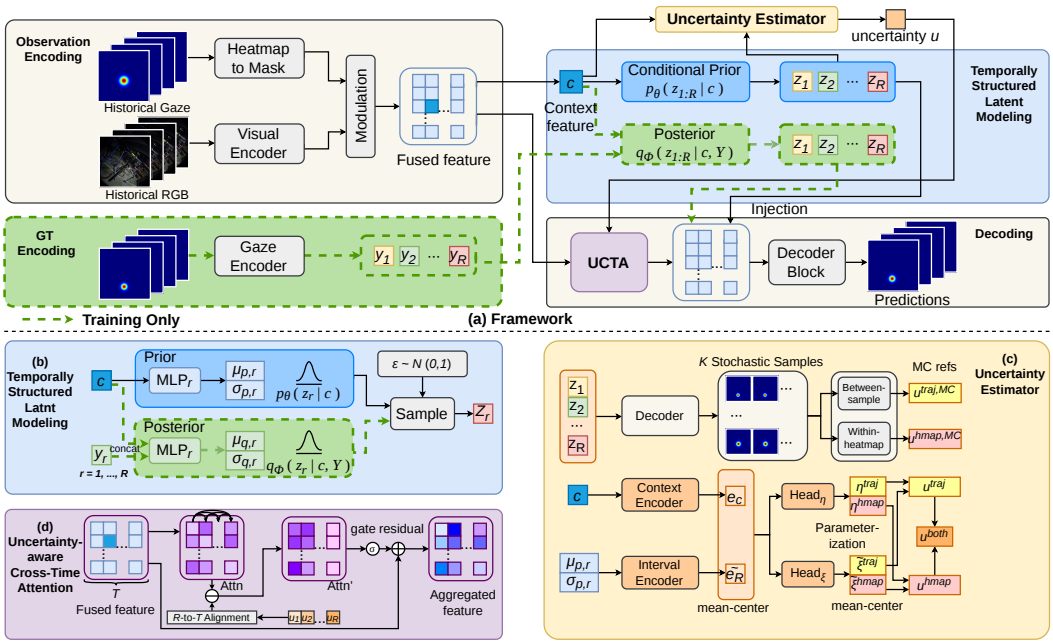


Figure 2: VANTAGE framework: (a) Model architecture, (b) Temporally Structured Variational Modeling, (c) Uncertainty Estimator, (d) Uncertainty-aware Cross-Time Attention (UCTA)

et al., 2025). In egocentric behavior modeling, SAGE (Kuang & Agarwal, 2026) examined gaze-conditioned action uncertainty using predictive entropy, while our focus is on uncertainty in future gaze itself. For spatial uncertainty quantification, discrete heatmap entropy is invariant to the spatial arrangement of probability values and does not directly measure the spatial dispersion. We instead study uncertainty across both predicted gaze trajectories and within individual heatmaps, with an emphasis on temporally structured uncertainty estimation and its relevance to forecast reliability.

3 Methodology

3.1 Task Formulation

We study **probabilistic egocentric gaze anticipation**. Let $V_h = \{v_\tau\}_{\tau=t-t_h}^{t}$ denote the observed history of t_h RGB frames where $v_\tau \in \mathbb{R}^{C\times H\times W}$, and $G_h = \{g_\tau\}_{\tau=t-t_h}^{t}$ denote the corresponding historical gaze heatmap sequence, where $g_\tau \in \mathbb{R}^{H\times W}$ is a Gaussian gaze heatmap. Following common practice (Lai et al., 2022; 2024b), each gaze heatmap is constructed from the observed gaze coordinates (x, y) using a Gaussian kernel. Let $G_f = \{g_\tau\}_{\tau=t+1}^{t+t_f}$ denote the future gaze heatmap sequence over the forecast horizon t_f. Conventional gaze anticipation predicts one deterministic future sequence: $\hat{G}_f = f_\theta(V_h, G_h)$. In contrast, we reformulate gaze anticipation as learning the conditional predictive distribution over plausible future gaze heatmap sequences:

$$p_\theta(G_f|V_h, G_h), \tag{1}$$

which is distinct from the Gaussian spatial representation of an individual gaze heatmap itself.

3.2 VANTAGE: Variational Anticipation with Temporal-Aware Gaze Uncertainty Estimation

Overview. We propose VANTAGE , a temporally structured conditional variational framework that unifies latent modeling, uncertainty estimation, and temporal attention over a shared interval-based horizon decomposition (Fig. 2(a)). Concretely, VANTAGE (i) decomposes the horizon into intervals and introduces temporally structured latent representations to model future gaze variability (Sec. 3.2; Fig. 2(b)), (ii) estimates trajectory- and heatmap-level uncertainty at the shared interval level before

decoding (Sec. 3.2; Fig. 2(c)), and (iii) exploits the uncertainty estimates to bias cross-time attention, down-weighting the relatively uncertain intervals (Sec. 3.2; Fig. 2(d)).

We leverage an MViT (Fan et al., 2021; Lai et al., 2024b) as the forecasting backbone, responsible for feature extraction and decoding. We fuse the historical gaze heatmaps with the encoded visual feature through modulation: $\widetilde{\mathbf{F}} = \mathbf{F} \odot (1 + \gamma\, \mathcal{A}(G_h))$, where $\mathbf{F}$ is the final-stage visual feature of MViT, $\mathcal{A}$ aligns gaze heatmaps with the feature grid and γ is a gate. The pooled fused feature, denoted by $c = \mathrm{Pool}(\widetilde{\mathbf{F}})$ is then used to condition probabilistic forecasting, while the MViT backbone remains otherwise unchanged. Implementation details are provided in Appendix A.1. VANTAGE follows a conditional variational formulation. We additionally adopt a lightweight discrete mode variable as an auxiliary design choice, with details provided in Appendix A.2.

Temporally Structured Conditional Variational Modeling. We decompose the forecast horizon into R contiguous intervals of length $\Delta t = t_f/R$, such that $G_f^{(r)} = \{g_\tau\}_{\tau=t+1+(r-1)\Delta t}^{t+r\Delta t}$, where $r = 1, \ldots, R$, and associate each interval with a continuous latent $z_r \in \mathbb{R}^{d_z}$. The interval-wise conditional prior and approximate posterior factorize as:

$$p_\theta(z_{1:R} \mid c) = \prod_{r=1}^{R} \mathcal{N}\big(z_r | \mu_r^p, \mathrm{diag}((\sigma_r^p)^2)\big)\,, \tag{2}$$

$$q_\phi(z_{1:R} \mid c, G_f) = \prod_{r=1}^{R} \mathcal{N}\big(z_r | \mu_r^q, \mathrm{diag}((\sigma_r^q)^2)\big)\,, \tag{3}$$

where $[\mu_r^p, \log((\sigma_r^p)^2)] = \mathrm{MLP}_\theta^{(r)}(c)$ and $[\mu_r^q, \log((\sigma_r^q)^2)] = \mathrm{MLP}_\phi^{(r)}([c; \mathbf{y}_r])$, with $\mathbf{y}_r$ mapped from $G_f^{(r)}$ by a shared future encoder during training. We sample interval latents from the approximate posterior during training and from the conditional prior at inference. Before decoding, each sampled latent is smoothly bounded $\tilde{z}_r = b\tanh(z_r/b)$, where b controls the range.

At decoder stage s, the temporal feature grid is decomposed using the same rule, and each latent is injected into its corresponding region: $\mathbf{H}^{(s,r)} \leftarrow \mathbf{H}^{(s,r)} + \gamma_{s,r}^z W_{s,r} \tilde{z}_r$, where $W_{s,r}$ is a projection layer and $\gamma_{s,r}^z$ is a learnable gate. This interval-specific injection associates distinct latents with temporal regions corresponding to distinct intervals of the forecast horizon, enabling the variability, while the shared decoder preserves cross-interval interaction. Marginalizing the latents induces the predictive distribution $p_\theta(G_f \mid c)$, whose marginal over interval r is

$$p_\theta(G_f^{(r)} \mid c) = \int p_\theta(G_f^{(r)} \mid z_{1:R}, c)\, p_\theta(z_{1:R} \mid c)\, dz_{1:R}. \tag{4}$$

Although the interval-specific latents follow a factorized prior, the shared decoder jointly processes the full forecast horizon, therefore, the decoded future sequence is not factorized across intervals.

Output-Space Predictive Uncertainty Estimation. We characterize two complementary forms of predictive uncertainty defined in the gaze output space: (i) trajectory-level uncertainty, which describes the between-trajectory variability in predicted gaze centers, and (ii) heatmap-level uncertainty, which describes spatial dispersion within an individual predicted gaze heatmap. Importantly, latent prior variance does not directly quantify either form in output space, since latent variables are transformed nonlinearly by the decoder. We therefore estimate both forms of interval-wise predictive uncertainty from stochastic predictions in the gaze output space.

For K stochastic predictions sampled from the conditional prior, let $\hat{g}_{k,\tau}$ denote the spatially normalized predicted gaze heatmap of sample k at future time point τ, with gaze center $(x, y)_{k,\tau} = \mathrm{SoftArgmax}(\hat{g}_{k,\tau})$. We further denote the mean of the predicted heatmaps as $\bar{\hat{g}}_\tau = \frac{1}{K}\sum_{k=1}^{K} \hat{g}_{k,\tau}$, and the mean gaze center as $(\bar{x}, \bar{y})_\tau = \frac{1}{K}\sum_{k=1}^{K}(x, y)_{k,\tau}$. We first define the per-time-point MC uncertainty forms as:

$$u_\tau^{\mathrm{traj,MC}} = \left[\frac{1}{K}\sum_{k=1}^{K} \|(x, y)_{k,\tau} - (\bar{x}, \bar{y})_\tau\|_2^2\right]^{1/2}, \tag{5}$$

$$u_\tau^{\mathrm{hmap,MC}} = \left[\frac{1}{K}\sum_{k=1}^{K}\sum_{i=0}^{W-1}\sum_{j=0}^{H-1} \hat{g}_{k,\tau}(i, j) \left\| \left(\frac{i}{W-1}, \frac{j}{H-1}\right) - (x, y)_{k,\tau} \right\|_2^2\right]^{1/2}. \tag{6}$$

Before square rooting and numerical flooring, the two quantities satisfy the law-of-total-variance:

$$\sum_{i=0}^{W-1}\sum_{j=0}^{H-1}\bar{\hat{g}}_\tau(i,j)\left\|\left(\frac{i}{W-1},\frac{j}{H-1}\right)-(\bar{x},\bar{y})_\tau\right\|_2^2=\left(u_\tau^{\mathrm{traj,MC}}\right)^2+\left(u_\tau^{\mathrm{hmap,MC}}\right)^2. \tag{7}$$

Where the first term corresponds to trajectory-level uncertainty, and the second term corresponds to heatmap-level uncertainty. Then the interval-wise predictive uncertainty is the average of the per-time-point quantities, $u_r^{\mathrm{form,MC}}=\frac{1}{\Delta t}\sum_\tau u_\tau^{\mathrm{form,MC}}$, where $\mathrm{form}\in\{\mathrm{traj},\mathrm{hmap}\}$.

We propose to exploit this interval-wise predictive uncertainty in the same prediction process, because it provides a reliability signal for subsequent temporal reasoning. However, direct MC uncertainty estimation is post-hoc, since it becomes available only after multiple stochastic futures have been decoded. We therefore learn an output-space estimator that predicts the interval-wise predictive uncertainty before decoding from the context feature c and interval-wise conditional-prior statistics $[\mu_r^p,\log((\sigma_r^p)^2)]$. We parameterize each form of interval-wise log-uncertainty by mean–deviation decomposition: $\log\hat{u}_r^{\mathrm{form}}=\eta^{\mathrm{form}}+\left(\xi_r^{\mathrm{form}}-\bar{\xi}^{\mathrm{form}}\right)$,where η^{form} captures the sample-level global log-uncertainty, and $\left(\xi_r^{\mathrm{form}}-\bar{\xi}^{\mathrm{form}}\right)$ denotes the relative deviation, $\bar{\xi}^{\mathrm{form}}=\frac{1}{R}\sum_{r=1}^{R}\xi_r^{\mathrm{form}}$. The estimator uses a global head and an interval-residual head to predict the overall uncertainty level and relative uncertainty deviation, respectively, for both uncertainty forms. After parameterization, we supervise trajectory-level and heatmap-level uncertainty estimates in log-space using their corresponding MC references. During MC target construction, uncertainty feedback is neutralized so that the reference targets are independent of the current uncertainty estimates. Model design details are provided in Appendix A.3.

Uncertainty-Aware Cross-Time Attention We exploit the estimated interval-wise predictive uncertainty as a reliability signal for the subsequent temporal reasoning. Our key idea is that temporal information associated with higher-uncertainty intervals should exert less influence on future prediction. To do this, we introduce *Uncertainty-Aware Cross-Time Attention* (UCTA), which performs attention along the temporal dimension of the feature grid to establish cross-time dependencies, and incorporates interval-wise uncertainty as an additive bias to down-weight the attention logits.

Let $\mathbf{H}^{(s)}\in\mathbb{R}^{T_s\times H_s\times W_s\times C_s}$ denote the feature representation at decoder stage s. For each spatial position p, UCTA operates on the temporal sequence $\mathbf{H}_p^{(s)}\in\mathbb{R}^{T_s\times C_s}$. For each attention head, the content attention between query time point τ and key time point τ' is $A_{p,\tau,\tau'}^{(s)}=\mathbf{Q}_{p,\tau}^{(s)}\mathbf{K}_{p,\tau'}^{(s)\top}/\sqrt{d}$, where d is the head dimension. At decoder stage s, we assign each temporal position the score of its corresponding interval and standardize the scores over time: $\tilde{\ell}_\tau^{(s)}=\frac{\ell_{r_s(\tau)}-\mu_\ell^{(s)}}{\sigma_\ell^{(s)}+\epsilon}$, where $\ell_r=\log\hat{u}_r^{\mathrm{form}}$, and $r_s(\tau)$ denotes the interval containing temporal position τ, and $\mu_\ell^{(s)}$ and $\sigma_\ell^{(s)}$ are the corresponding temporal mean and standard deviation. This normalization retains relative uncertainty across temporal positions while removing the absolute uncertainty scale. UCTA incorporates the resulting reliability signal along the key dimension:

$$\hat{A}_{p,\tau,\tau'}^{(s)}=A_{p,\tau,\tau'}^{(s)}-\tilde{\ell}_{\tau'}^{(s)}. \tag{8}$$

Consequently, the attention weights satisfy $\alpha_{p,\tau,\tau'}^{(s)}\propto\exp\left(A_{p,\tau,\tau'}^{(s)}\right)\exp\left(-\tilde{\ell}_{\tau'}^{(s)}\right)$, and the UCTA output is obtained by aggregating the value features using the uncertainty-modulated attention weights: $\mathbf{O}_{p,\tau}^{(s)}=\mathrm{Proj}^{(s)}\left(\sum_{\tau'}\alpha_{p,\tau,\tau'}^{(s)}\mathbf{V}_{p,\tau'}^{(s)}\right)$, At each decoder stage, UCTA is applied before the interval-specific latent injection, and its output is integrated through a stage-specific gated residual connection, $\mathbf{H}^{(s)}\leftarrow\mathbf{H}^{(s)}+\gamma_s^{\mathrm{ucta}}\mathbf{O}^{(s)}$. Where γ_s^{ucta} is a learnable gate. Overall, UCTA realizes uncertainty-aware temporal interaction by jointly accounting for content attention and interval-wise predictive uncertainty: content attention establishes dependencies across temporal positions, while uncertainty modulates their relative contribution. When uncertainty is constant across time, the uncertainty bias vanishes and UCTA naturally reduces to uncertainty-free cross-time attention. Detailed implementation details are provided in Appendix A.4.

3.3 Learning Objective

We train VANTAGE with the joint objective

$$\mathcal{L}=\mathcal{L}_{\mathrm{rec}}+\lambda_{\mathrm{KL}}\mathcal{L}_{\mathrm{kl}}+\lambda_{\mathrm{unc}}\mathcal{L}_{\mathrm{unc}}+\lambda_{\mathrm{div}}\mathcal{L}_{\mathrm{div}}, \tag{9}$$

where $\mathcal{L}_{\text{unc}}$ is the output-space uncertainty estimation objective defined in Sec. 3.2.

Heatmap reconstruction. The reconstruction loss $\mathcal{L}_{\text{rec}}$ averages $D_{\text{KL}}(G_{f,\tau} \| \hat{G}_{f,\tau})$ over valid future steps, where $G_{f,\tau}$ and $\hat{G}_{f,\tau}$ denote the target and spatially normalized predicted heatmaps, respectively. During training, we use posterior Best-of-K sampling and select the sample with the lowest reconstruction criterion for prediction.

Latent regularization. For the interval-specific latents, we minimize the average posterior–prior KL divergence $\mathcal{L}^z_{\text{kl}} = \frac{1}{R}\sum_{r=1}^{R} D_{\text{KL}}\left(q_\phi(z_r \mid c, G_f^{(r)}) \parallel p_\theta(z_r \mid c)\right)$. The full latent regularization additionally includes categorical KL and batch-level balancing terms for the auxiliary discrete mode: $\mathcal{L}_{\text{kl}} = \mathcal{L}^z_{\text{kl}} + \lambda_m \mathcal{L}^m_{\text{kl}} + \lambda_b \mathcal{L}_{\text{bal}}$. Free-bits regularization is applied to the continuous latent KL.

Diversity regularization. We encourage trajectories decoded from different latent samples to remain separated: $\mathcal{L}_{\text{div}} = \max\left(0, m_{\text{div}} - \frac{1}{t_f}\sum_\tau \|(x,y)_\tau - (x',y')_\tau\|_2\right)$, where $(x,y)_\tau$ and $(x',y')_\tau$ are gaze centers from any two different trajectories, and m_{div} is the diversity margin. Additional loss details are provided in Appendix A.5.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on three public egocentric datasets: Ego-Exo4D (Grauman et al., 2024), Ego4D (Grauman et al., 2022), and EGTEA Gaze+ (Li et al., 2021). We clip each video into non-overlapping 5-second clips, using the first 86 frames as historical observations and the remaining as future targets. We uniformly sample 8 frames for both sequences, and use 256×256 crops for RGB frames and gaze heatmaps.

Implementation details. The dimension of the latent representation is $d_z = 32$. Four-head UCTA is applied at all decoder stages. We train for 25 epochs with AdamW using a learning rate of 10^{-5}, batch size 8, KL warm-up, and free bits on a single 24GB NVIDIA RTX 6000. We use $K_{\text{MC}} = 8$ stochastic samples and temperature $Temp = 10$ for uncertainty supervision.

4.2 Evaluation Protocol

Predictive accuracy. To remain comparable with prior methods, we report deterministic metrics including F1, precision, and recall, from a single prediction. For probabilistic evaluation, we draw $N = 20$ stochastic predictions and recover gaze coordinates from the predicted heatmaps by soft-argmax. We report minADE@N, which measures the displacement error of the best-matching sampled trajectory, and meanADE@N, which measures the mean error across all sampled trajectories.

Distribution quality. We evaluate the predicted gaze distribution using Negative Log-Likelihood (NLL) and Continuous Ranked Probability Score (CRPS), both computed from N stochastic predictions in pixel space (crop size). NLL measures the density assigned to the observed gaze, and CRPS jointly assesses predictive location and spread. Detailed definitions are provided in Appendix B.

Uncertainty reliability. We evaluate uncertainty from: MC-reference fidelity, realized-error association, and selective reliability. MC-reference fidelity measures Spearman's rank correlation coefficient (Spearman's ρ) to assess whether the uncertainty estimates can faithfully recover their corresponding output-space MC references. Realized-error association assesses whether higher uncertainty identifies predictions with larger realized forecasting error using interval-stratified uncertainty quantiles to control for differences across forecast intervals. Selective reliability assesses whether uncertainty provides a useful ranking of prediction reliability, quantified through risk–coverage analysis and AURC. Confidence intervals (CI) are computed using take-level cluster bootstrap.

4.3 Main Results

Probabilistic Performance. Tab. 1 evaluates the probabilistic performance of the proposed models. MC DropPath provides a generic stochastic baseline. It remains clearly inferior in metrics, showing that backbone stochasticity alone is insufficient to model the variability of future gaze. Our mod-

Table 1: Distributional metrics with $N = 20$ stochastic predictions. All models share MViT backbone. MC DropPath (Huang et al., 2016; Müller et al., 2026) retains training-time stochastic depth at inference; Global CVAE uses a single global latent, while TS-CVAE (Temporally Structured CVAE) uses $R = 4$ interval-specific latents without UCTA. V and G denote video and historical gaze.

		Ego-Exo4D				Ego4D				EGTEA Gaze+			
Method	Input	minADE @20 ↓	meanADE @20 ↓	NLL ↓	CRPS ↓	minADE @20 ↓	meanADE @20 ↓	NLL ↓	CRPS ↓	minADE @20 ↓	meanADE @20 ↓	NLL ↓	CRPS ↓
MC DropPath	V	25.78	30.03	129.99	34.24	41.39	44.46	580.15	54.41	52.01	54.07	2404.04	74.00
Global CVAE	V	22.61	**26.95**	49.65	**29.36**	45.59	47.17	2915.06	57.57	45.95	54.29	575.62	69.26
VANTAGE	V	**22.56**	28.84	**32.88**	30.21	**33.40**	**43.91**	**22.25**	**43.45**	**37.02**	**48.47**	**328.19**	**58.38**
MC DropPath	V+G	26.85	34.40	44.92	36.76	31.73	35.90	253.80	41.66	40.91	46.97	349.91	60.08
Global CVAE	V+G	17.98	**20.74**	101.15	22.66	30.99	**32.06**	3513.67	38.64	41.91	47.88	1252.17	61.46
TS-CVAE	V+G	20.18	43.97	**9.56**	30.75	29.91	47.13	**17.85**	43.64	33.74	49.80	**24.43**	57.54
VANTAGE	V+G	**16.58**	23.65	10.74	**19.91**	**25.66**	33.19	34.82	**32.24**	**31.25**	**41.41**	69.20	**46.73**

Table 2: Comparison with existing deterministic methods. We report single-F1, precision, and recall for our models. A denotes audio, and A is not applicable to the EGTEA Gaze+ dataset.

			Ego-Exo4D			Ego4D			EGTEA Gaze+		
Method	Backbone	Input	F1 ↑	Prec. ↑	Rec. ↑	F1 ↑	Prec. ↑	Rec. ↑	F1 ↑	Prec. ↑	Rec. ↑
MViT (Fan et al., 2021)	MViT	V	0.497	0.483	0.511	0.326	0.246	0.486	0.278	0.192	0.501
MViT (Fan et al., 2021)	MViT	V+G	0.607	0.578	**0.639**	**0.532**	0.461	**0.629**	**0.474**	0.389	**0.608**
GLC (Lai et al., 2024a)	MViT	V	0.514	0.477	0.556	0.364	0.281	0.517	0.290	0.203	0.512
CSTS (Lai et al., 2024b)	MViT	V+A	0.492	0.478	0.507	0.369	0.285	0.524	—	—	—
VANTAGE (Ours)	MViT	V	0.448	0.458	0.441	0.317	0.431	0.255	0.231	0.415	0.163
VANTAGE (Ours)	MViT	V+G	**0.614**	**0.621**	0.608	0.518	**0.596**	0.459	0.444	**0.541**	0.379

els improve different aspects of probabilistic forecasting. TS-CVAE achieves the lowest NLL on all datasets in V+G setting, indicating that temporally structured latents assign higher probability density to the realized gaze. The high NLL of Global CVAE is consistent with overly concentrated predictions (minADE/meanADE) that deviate from the realized gaze. VANTAGE further achieves the best minADE and CRPS on all datasets while substantially reducing meanADE relative to TS-CVAE. These results show that temporal structure improves the distribution of stochastic predictions, particularly sample-based likelihood, while UCTA further improves accuracy and quality.

Comparison with existing methods. Tab. 2 compares the deterministic performance of VANTAGE with existing methods. On Ego-Exo4D, VANTAGE with historical gaze achieves the best F1 score of 0.614, outperforming the MViT baseline with the same backbone and input (0.607). In other cases, VANTAGE's single F1 is slightly lower but still competitive, reflecting the sampling variability introduced by probabilistic modeling. This shows that probabilistic modeling of multiple plausible futures entails only a modest trade-off in single-prediction accuracy. The gain from gaze modality further highlights its complementary value. Figure 1 provides a qualitative comparison.

4.4 Ablation Studies

Temporally Structured Modeling. Tab. 3 evaluates the effect of temporal latent granularity by setting different numbers of intervals R. Temporally structured latent variables improve both best-of-N accuracy and distributional quality. Compared with the single global latent ($R = 1$), $R = 2, 4$ reduce minADE, NLL, and CRPS, with $R = 4$ achieving the best performance on all three metrics. Importantly, the capacity-matched $R = 1$ variant with four MLPs remains inferior to $R = 4$ on minADE, NLL, and CRPS, indicating that the gain comes from temporally structured modeling rather than additional model capacity. Together, these results support our core idea that future gaze uncertainty varies over the forecast horizon and benefits from temporally structured latents rather than globally modeled latents. We therefore use $R = 4$ in all subsequent experiments.

Uncertainty-Aware Cross-Time Attention. Table 4 evaluates different uncertainty signals for guiding UCTA. Pred-Traj achieves the best minADE (16.58) and CRPS (19.91), while remaining essentially tied on meanADE. In contrast, Raw-σ is generally weaker than the uncertainty estimates,

Table 3: Ablation of number of intervals in VANTAGE on Ego-Exo4D. The $R = 1$, 4-MLP variant is capacity matched to $R = 4$.

R	# MLPs	minADE @20 ↓	meanADE @20 ↓	NLL ↓	CRPS ↓
1	1	17.59	23.92	22.86	20.79
1	4	17.96	**22.73**	16.95	20.89
2	2	16.61	24.13	12.47	20.13
3	3	16.69	24.00	27.03	20.46
4	4	**16.58**	23.65	**10.74**	**19.91**

Table 4: Ablation of the uncertainty signal used for UCTA on Ego-Exo4D. CTA refers to cross-time attention.

Signal	minADE @20 ↓	meanADE @20 ↓	NLL ↓	CRPS ↓
None (CTA)	17.06	25.08	**10.10**	20.29
Raw-σ	17.06	25.32	10.24	20.29
Pred-Traj.	**16.58**	23.65	10.74	**19.91**
Pred-Hmap.	16.81	24.01	10.66	19.97
Pred-Both	16.84	**23.62**	11.16	20.09

indicating that latent prior scale alone is a less effective signal for UCTA. Pred-Hmap and Pred-Both also do not improve upon Pred-Traj on minADE or CRPS, indicating that different forms of predictive uncertainty are not equally useful as uncertainty signals. Overall, these results show that exploiting predictive uncertainty can improve future gaze distribution prediction, and trajectory-level uncertainty provides the most effective signal for UCTA. They also support explicitly decomposing and predicting distinct uncertainty forms, rather than a single undifferentiated form. We therefore use trajectory-level uncertainty in the final VANTAGE, and further analyze the distinct roles of trajectory- and heatmap-level uncertainty in Sec. 4.5.

4.5 Model Analysis

Predictive Uncertainty Estimation. Table 5 evaluates the fidelity of the uncertainty estimates to their corresponding output-space MC references, with Raw-σ as a latent-space baseline. Both trajectory- and heatmap-level uncertainty estimates consistently achieve higher Spearman's ρ. For trajectory-level uncertainty, the overall correlation improves from 0.531 to 0.632; for heatmap uncertainty, it increases from -0.381 to 0.575. These results show that latent prior scale alone is incomplete to reflect predictive uncertainty in the output space. On the other hand, the advantage is more pronounced in temporal fidelity. Pred-Traj increases the temporal correlation from 0.292 to 0.619, while Pred-Hmap improves it from 0.271 to 0.671. The learned estimator recovers how predictive uncertainty varies over the forecast horizon, and also supports modeling future uncertainty as temporally structured rather than globally shared.

Table 5: Spearman's ρ of predictive uncertainty estimates to $K = 20$ output-space MC references.

MC Ref.	Signal	Overall ρ ↑	Sample ρ ↑	Temporal ρ ↑
MC-Traj.	Raw-σ	0.531	0.284	0.292
	Pred-Traj.	**0.632**	**0.372**	**0.619**
MC-Hmap.	Raw-σ	-0.381	-0.063	0.271
	Pred-Hmap.	**0.575**	**0.180**	**0.671**

Realized-Error Reliability. As shown in the Appendix Fig. 4(a), realized ADE increases monotonically across interval-stratified Pred-Traj uncertainty quintiles $(18.58, 19.07, 19.86, 21.20, 24.08)$, while not for Raw-$\sigma$. Because the quantiles are formed separately within each forecast interval, this trend shows that Pred-Traj discriminates forecast reliability beyond the effect of prediction horizon. This is further supported by selective prediction. As shown in Fig. 3(a), Pred-Traj achieves the lowest AURC among the evaluated uncertainty signals (16.18), compared with Raw-σ (16.63), Pred-Both (16.94), and Pred-Hmap (24.10); random ranking and the error oracle yield 20.56 and 10.04, respectively. Its improvement over Raw-σ is also robust under paired take-level bootstrap ($\Delta\mathrm{AURC} = -0.45$, 95% CI $[-0.63, -0.28]$). These results establish trajectory-level uncertainty as the most effective evaluated signal for ranking forecast reliability. The contrasting behavior of the uncertainty forms further indicates that they are not interchangeable as reliability signals, motivating the semantic analysis that follows.

Uncertainty Semantics. We isolate the effect of spatial decoding by varying the spatial-softmax temperature while keeping the model, stochastic decoder outputs, and realized errors fixed. The figures show that MC heatmap dispersion increases from 0.098 at $Temp = 5$ to 0.330 at $Temp = 20$ (Fig. 4(b)), while its correlation with realized ADE changes from 0.097 to -0.276 (Fig. 3(b)). In contrast, trajectory dispersion remains positively associated with ADE throughout the sweep ($\rho = 0.199$–0.412). Heatmap dispersion is instead strongly correlated with spatial entropy ($\rho =$

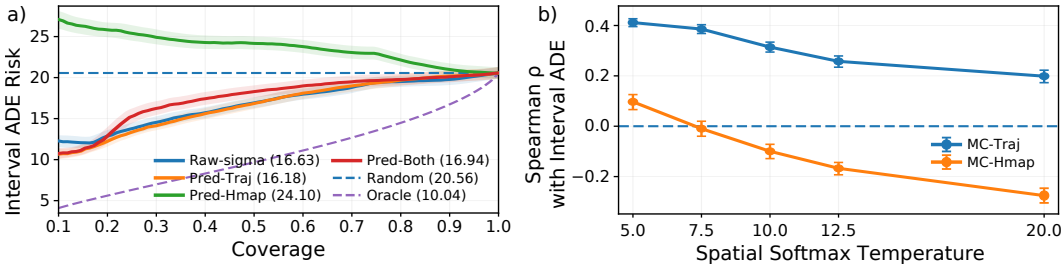


Figure 3: Uncertainty analysis on Ego-Exo4D. (a) Selective risk–coverage across uncertainty signals. (b) Error correlation of trajectory- and heatmap-level dispersion across spatial-softmax temperatures. Shaded regions denote 95% take-level bootstrap confidence intervals.

0.68–0.84), indicating that it primarily characterizes within-heatmap spatial spread rather than a stable measure of forecast reliability. This distinction is not attributable to estimator failure: Table 5 shows that Pred-Hmap remains well aligned with its MC target ($\rho = 0.575$). Together, these results show that the two uncertainty forms capture different aspects of predictive ambiguity: heatmap-level uncertainty describes spatial dispersion within a prediction, whereas trajectory-level uncertainty is more consistently associated with realized forecast reliability. This supports decomposing predictive uncertainty according to its semantic role rather than treating different forms as interchangeable confidence signals.

5 Conclusion

We studied egocentric gaze anticipation as conditional probabilistic forecasting, motivated by the observation that future gaze is inherently uncertain and its conditional variability is temporally structured. We introduced VANTAGE, which unifies temporally structured latent representation, output-space predictive uncertainty estimation, and uncertainty-aware cross-time attention over a shared interval-wise decomposition. By decomposing interval-wise predictive uncertainty into between-trajectory and within-heatmap forms and estimating them, the uncertainty is available before decoding and the trajectory-level uncertainty can serve as a reliability signal for temporal reasoning.

Experiments on Ego-Exo4D (Grauman et al., 2024), Ego4D (Grauman et al., 2022), and EGTEA Gaze+ (Li et al., 2021) datasets demonstrate that VANTAGE entails only a modest trade-off in single-prediction accuracy relative to existing deterministic methods (Lai et al., 2022; 2024b), while achieving the best F1 on Ego-Exo4D with historical gaze. For distributional performance, VANTAGE consistently outperforms MC DropPath (Huang et al., 2016; Müller et al., 2026), a generic stochastic baseline, across predictive accuracy and distributional metrics. The result shows that temporally structured latent representations improve distributional modeling over a global latent representation, while UCTA improves the quality of sampled future predictions. Our uncertainty analysis further shows that latent prior scale alone is an incomplete proxy for uncertainty in the gaze output space: the learned estimator more faithfully recovers both the trajectory- and heatmap-level uncertainties with temporal structures than Raw-σ. Moreover, predictive dispersion is heterogeneous in its semantic role. Trajectory-level uncertainty is associated with realized forecasting error, supports better selective prediction, and provides the most effective signal for UCTA. In contrast, heatmap-level uncertainty captures within-heatmap spatial dispersion and is strongly affected by the spatial decoding scale, making it a meaningful measure of predictive spread but not a stable proxy for forecast reliability. These results suggest that different forms of predictive uncertainty should be distinguished by their semantic roles rather than collapsed into a single uncertainty signal.

Several directions remain for future work. The current interval decomposition is fixed and uniform, and our experiments focus on an MViT-family backbone. Extending VANTAGE to learned non-uniform temporal structures and broader architectures, forecast horizons, and downstream decision-making settings would further examine the generality of the proposed uncertainty formulation.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for language editing, polishing, improving the clarity and organization of the narrative, debugging and AI-assisted code implementation. We have not used generative AI tools for research ideation, retrieval, generating synthetic datasets, generating experimental data or results, and other required disclosure tasks are not applicable to this work. Additionally, we didn't use generative AI tools for tasks with recommended disclosure. We have reviewed all AI-assisted work, LLM-generated code was verified and tested for correctness by 2 authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Inhwan Bae, Young-Jae Park, and Hae-Gon Jeon. Singulartrajectory: Universal trajectory predictor using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17890–17901, 2024.

Wentao Bao, Lele Chen, Libing Zeng, Zhong Li, Yi Xu, Junsong Yuan, and Yu Kong. Uncertainty-aware state space transformer for egocentric 3d hand trajectory forecasting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 13702–13711, 2023.

Anna Belardinelli. Gaze-based intention estimation: principles, methodologies, and applications in hri. *ACM Transactions on Human-Robot Interaction*, 13(3):1–30, 2024.

Guillem Capellera, Antonio Rubio, Luis Ferraz, and Antonio Agudo. Unified uncertainty-aware diffusion for multi-agent trajectory modeling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22476–22486, 2025.

Giuseppe Cartella, Vittorio Cuculo, Alessandro D'Amelio, Marcella Cornia, Giuseppe Boccignone, and Rita Cucchiara. Modeling human gaze behavior with diffusion models for unified scanpath prediction. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16206–16216. IEEE, 2025.

Patrick Dendorfer, Sven Elflein, and Laura Leal-Taixé. Mg-gan: A multi-generator model preventing out-of-distribution samples in pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 13158–13167, 2021.

Michael Dorr, Thomas Martinetz, Karl R Gegenfurtner, and Erhardt Barth. Variability of eye movements when viewing dynamic natural scenes. *Journal of vision*, 10(10):28–28, 2010.

Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6824–6835, 2021.

Alexander Goettker, Nathaniel Powell, and Mary Hayhoe. Approaches to understanding natural behavior. *Journal of Vision*, 25(6):12–12, 2025.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19383–19400, 2024.

Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2255–2264, 2018.

Mary M Hayhoe and Jonathan Samir Matthis. Control of gaze in natural environments: effects of rewards and costs, uncertainty and memory in target selection. *Interface focus*, 8(4):20180009, 2018.

Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *European conference on computer vision*, pp. 646–661. Springer, 2016.

Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 754–769, 2018.

Yifei Huang, Minjie Cai, Zhenqiang Li, Feng Lu, and Yoichi Sato. Mutual context network for jointly estimating egocentric gaze and action. *IEEE Transactions on Image Processing*, 29:7795–7806, 2020.

Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian Reid, Hamid Rezatofighi, and Silvio Savarese. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. *Advances in neural information processing systems*, 32, 2019.

Chenyi Kuang and Nakul Agarwal. Sage: Synchronized action-gaze recognition and anticipation for human behavior understanding. *arXiv preprint arXiv:2607.04017*, 2026.

Bolin Lai, Miao Liu, Fiona Ryan, and James M Rehg. In the eye of transformer: Global-local correlation for egocentric gaze estimation. In *33rd British Machine Vision Conference Proceedings, BMVC 2022*, 2022.

Bolin Lai, Miao Liu, Fiona Ryan, and James M Rehg. In the eye of transformer: Global–local correlation for egocentric gaze estimation and beyond. *International Journal of Computer Vision*, 132(3):854–871, 2024a.

Bolin Lai, Fiona Ryan, Wenqi Jia, Miao Liu, and James M Rehg. Listen to look into the future: Audio-visual egocentric gaze anticipation. In *European Conference on Computer Vision*, pp. 192–210. Springer, 2024b.

Mihee Lee, Samuel S Sohn, Seonghyeon Moon, Sejong Yoon, Mubbasir Kapadia, and Vladimir Pavlovic. Muse-vae: Multi-scale vae for environment-aware long term trajectory prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2221–2230, 2022.

Chengyang Li, Kaiyi Xiong, Yuan Xu, Lei Qian, Yizhou Wang, and Wentao Zhu. Gazevla: Learning human intention for robotic manipulation. *arXiv preprint arXiv:2604.22615*, 2026.

Guopeng Li, Zirui Li, Victor L Knoop, and Hans van Lint. Unravelling uncertainty in trajectory prediction using a non-parametric approach. *Transportation Research Part C: Emerging Technologies*, 163:104659, 2024.

Yin Li, Alireza Fathi, and James M Rehg. Learning to predict gaze in egocentric video. In *Proceedings of the IEEE international conference on computer vision*, pp. 3216–3223, 2013.

Yin Li, Miao Liu, and James M Rehg. In the eye of the beholder: Gaze and actions in first person video. *IEEE transactions on pattern analysis and machine intelligence*, 45(6):6731–6747, 2021.

Yujie Li, Xinghe Wang, Zihang Ma, Yifu Wang, and Michael C Meyer. Swingaze: Egocentric gaze estimation with video swin transformer. In *2023 IEEE 16th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC)*, pp. 123–127. IEEE, 2023.

Yao Liu, Zesheng Ye, Rui Wang, Binghao Li, Quan Z Sheng, and Lina Yao. Uncertainty-aware pedestrian trajectory prediction via distributional diffusion. *Knowledge-Based Systems*, 296: 111862, 2024.

Charles Marx, Volodymyr Kuleshov, and Stefano Ermon. Calibrated probabilistic forecasts for arbitrary sequences. *Transactions on machine learning research*, 2025:3807, 2025.

Adam T Müller, Tobias Rögelein, and Nicolaj C Stache. Monte carlo stochastic depth for uncertainty estimation in deep learning. *arXiv preprint arXiv:2604.12719*, 2026.

Heqian Qiu, Lanxiao Wang, Taijin Zhao, Zhaofeng Shi, Xiang Li, Linfeng Xu, and Hongliang Li. Ego-pmove: Prompt-aware mixture of view experts network for egocentric gaze prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8565–8573, 2026.

Fiona Ryan, Ishwarya Ananthabhotla, Yijun Qian, Judy Hoffman, James M Rehg, Vamsi Krishna Ithapu, and Calvin Murdock. Forecasting 3d scanpaths in egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42824–42835, 2026.

Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *European conference on computer vision*, pp. 683–700. Springer, 2020.

Ali Shafti, Pavel Orlov, and A Aldo Faisal. Gaze-based, context-aware robotic system for assisted reaching and grasping. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 863–869. IEEE, 2019.

Yujia Wang, Fang-Lue Zhang, and Neil A Dodgson. Scantd: 360 scanpath prediction based on time-series diffusion. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 7764–7773, 2024.

Zichen Wang, Hao Miao, Senzhang Wang, Renzhi Wang, Jianxin Wang, and Jian Zhang. C2f-tp: A coarse-to-fine denoising framework for uncertainty-aware trajectory prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 12810–12817, 2025.

Brian Yang, Huangyuan Su, Nikolaos Gkanatsios, Tsung-Wei Ke, Ayush Jain, Jeff Schneider, and Katerina Fragkiadaki. Diffusion-es: Gradient-free planning with diffusion for autonomous driving and zero-shot instruction following. *arXiv preprint arXiv:2402.06559*, 2024.

Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris M Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9813–9823, 2021.

Heeseung Yun, Joonil Na, Jaeyeon Kim, Calvin Murdock, and Gunhee Kim. Gaze beyond the frame: Forecasting egocentric 3d visual span. *arXiv preprint arXiv:2511.18470*, 2025.

Mengmi Zhang, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Jiashi Feng. Anticipating where people will look using adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1783–1796, 2018.

Yuanshao Zhu, Yongchao Ye, Shiyao Zhang, Xiangyu Zhao, and James Yu. Difftraj: Generating gps trajectory with diffusion probabilistic model. *Advances in Neural Information Processing Systems*, 36:65168–65188, 2023.

A Additional Implementation Details

A.1 Observation Encoding and Gaze Fusion

VANTAGE uses the same MViT encoder–decoder backbone as the models in the main comparison. Let $\mathbf{F} \in \mathbb{R}^{T_L \times H_L \times W_L \times C}$ denote the final visual feature grid after removing the class token. Historical gaze is represented by normalized gaze coordinates together with the corresponding gaze type. For a historical gaze point $(x_\tau, y_\tau) \in [0,1]^2$, we first map the coordinates to the spatial resolution of the final feature grid:

$$\tilde{x}_\tau = x_\tau(W-1), \qquad \tilde{y}_\tau = y_\tau(H-1).$$

We then construct a type-aware Gaussian gaze mask

$$M_\tau(i,j) = a_\tau \exp\left(-\frac{(j-\tilde{x}_\tau)^2 + (i-\tilde{y}_\tau)^2}{2\sigma_\tau^2}\right), \tag{10}$$

where σ_τ and a_τ are determined by the gaze type. Specifically, fixation uses $(\sigma_\tau, a_\tau) = (1.0, 1.0)$, saccade uses $(2.5, 0.6)$, and invalid gaze uses $(1.5, 0)$; the default setting is $(1.5, 0.8)$.

The mask is constructed directly at the spatial resolution $H \times W$ of the final feature grid. When the number of historical gaze samples differs from the temporal resolution T of the visual tokens, the gaze masks are linearly interpolated along the temporal dimension. Each temporally aligned mask is subsequently normalized by its spatial maximum:

$$\bar{M}_\tau = \frac{M_\tau}{\max_{i,j} M_\tau(i,j) + \epsilon}. \tag{11}$$

Invalid gaze therefore produces a zero mask.

The resulting spatio-temporal gaze mask $\mathcal{A}(G_h) \in \mathbb{R}^{T \times H \times W}$ is flattened according to the visual-token ordering and applied through multiplicative modulation:

$$\widetilde{\mathbf{F}} = \mathbf{F} \odot (1 + \gamma\, \mathcal{A}(G_h)). \tag{12}$$

Thus, gaze-aligned locations receive larger feature responses while visual features outside the gaze region are preserved. The fused visual tokens are then averaged over the spatio-temporal grid,

$$c = \mathrm{Pool}(\widetilde{\mathbf{F}}), \tag{13}$$

and the resulting context feature c conditions the latent prior and uncertainty estimator.

UCTA is applied at all four decoder stages, whose feature dimensions are $\{768, 768, 384, 192\}$, using four attention heads.

A.2 Temporally Structured Latent Modeling

Continuous latents. For interval r, the conditional-prior and posterior parameters are

$$[\mu_r^p, \log((\sigma_r^p)^2)] = \mathrm{MLP}_\theta^{(r)}(c), \qquad [\mu_r^q, \log((\sigma_r^q)^2)] = \mathrm{MLP}_\phi^{(r)}([c; \mathbf{y}_r]), \tag{14}$$

where the shared future encoder maps $G_f^{(r)}$ to $\mathbf{y}_r$. Sampling is performed as

$$z_r = \begin{cases} \mu_r^q + \sigma_r^q \odot \epsilon_r, & \text{during training,} \\ \mu_r^p + \sigma_r^p \odot \epsilon_r, & \text{at inference,} \end{cases} \qquad \epsilon_r \sim \mathcal{N}(0, I). \tag{15}$$

Before decoder conditioning, we apply the bounded transformation

$$\tilde{z}_r = b \tanh(z_r / b), \qquad b = 4. \tag{16}$$

The same proportional interval partition is used to align future intervals with decoder temporal regions. At decoder stage s, interval-specific latent conditioning is applied as

$$\mathbf{H}^{(s,r)} \leftarrow \mathbf{H}^{(s,r)} + \gamma_{s,r}^z W_{s,r} \tilde{z}_r, \tag{17}$$

where $W_{s,r}$ is the interval-specific projection and $\gamma^z_{s,r}$ is a learnable latent-injection gate.

Auxiliary discrete mode. We additionally use a shared categorical mode $m \in \{1, \dots, M\}$, with $M = 4$, conditional prior $p_\theta(m \mid c)$ and posterior $q_\phi(m \mid c, G_f)$. The sampled mode embedding is shared across future intervals and enters decoder conditioning together with the interval-specific continuous latents. This auxiliary variable captures coarse future modes, while temporal variation is modeled by the interval-specific continuous latents.

During training, we employ posterior Best-of-K sampling to improve the selection of stochastic future realizations. For each training example, we draw K candidate realizations from the posterior, including a categorical mode sample and the corresponding continuous latent samples, decode each candidate, and select the one with the lowest heatmap reconstruction criterion. Only the selected realization is decoded with gradients for the prediction loss. This Best-of-K procedure is used only during training and does not alter the probabilistic sampling procedure at inference.

The categorical posterior–prior regularization is

$$\mathcal{L}^m_{\mathrm{kl}} = D_{\mathrm{KL}}\left(q_\phi(m \mid c, G_f) \,\|\, p_\theta(m \mid c)\right). \tag{18}$$

Let

$$\bar{q}(m) = \frac{1}{B}\sum_{n=1}^{B} q_\phi(m \mid c_n, G_{f,n})$$

denote the batch-averaged posterior distribution. We use

$$\mathcal{L}_{\mathrm{bal}} = D_{\mathrm{KL}}\left(\bar{q}(m) \,\|\, \mathrm{Unif}(M)\right) \tag{19}$$

to discourage collapse to a small subset of modes. We use a 16-dimensional mode embedding.

Latent-conditioned output residual. To facilitate stochastic information propagation to the output layer, we additionally map the concatenated latent conditioning to a low-resolution spatio-temporal residual:

$$\mathbf{z}_{\mathrm{cat}} = [\tilde{z}_1; \dots; \tilde{z}_R], \qquad \mathbf{B}_z = \mathrm{Interp}\left(f_{\mathrm{sp}}(\mathbf{z}_{\mathrm{cat}})\right), \tag{20}$$

and add the interpolated residual to the output logits,

$$\mathbf{L} = \mathbf{L}_{\mathrm{base}} + \lambda_z \mathbf{B}_z. \tag{21}$$

This supporting decoder design is kept fixed across the temporal-structure and UCTA ablations.

A.3 Uncertainty Estimator and MC Supervision

Estimator inputs. The uncertainty estimator uses the conditional-prior statistics already defined in the main text. For each interval r, the prior mean and log variance are encoded as

$$\mathbf{e}_r = E_{\mathrm{intv}}\left([\mu^p_r;\, \log((\sigma^p_r)^2)]\right), \qquad \mathbf{e}_c = E_{\mathrm{ctx}}(c). \tag{22}$$

We mean-center the interval embeddings:

$$\tilde{\mathbf{e}}_r = \mathbf{e}_r - \frac{1}{R}\sum_{j=1}^{R} \mathbf{e}_j. \tag{23}$$

The global uncertainty head predicts the overall log-uncertainty for the two uncertainty forms:

$$(\eta^{\mathrm{traj}}, \eta^{\mathrm{hmap}}) = \mathrm{MLP}_\eta\left(\left[\mathbf{e}_c;\, \frac{1}{R}\sum_{j=1}^{R}\mathbf{e}_j\right]\right). \tag{24}$$

The interval-residual head produces interval-wise scores:

$$(\xi^{\mathrm{traj}}_r, \xi^{\mathrm{hmap}}_r) = \mathrm{MLP}_\xi\left([\tilde{\mathbf{e}}_r;\, \mathbf{e}_c;\, \tilde{\mathbf{e}}_r \odot \mathbf{e}_c;\, |\tilde{\mathbf{e}}_r - \mathbf{e}_c|]\right). \tag{25}$$

Following the mean–deviation parameterization in the main text,

$$\log \hat{u}^{\mathrm{form}}_r = \eta^{\mathrm{form}} + \left(\xi^{\mathrm{form}}_r - \bar{\xi}^{\mathrm{form}}\right), \qquad \bar{\xi}^{\mathrm{form}} = \frac{1}{R}\sum_{r=1}^{R}\xi^{\mathrm{form}}_r, \tag{26}$$

where $\text{form} \in \{\text{traj}, \text{hmap}\}$. The estimator uses hidden dimension 64 and no dropout.

Monte Carlo reference construction. The output-space MC references are constructed from $K_{\text{MC}} = 8$ stochastic prior predictions using the trajectory- and heatmap-level definitions in Sec. 3.2. We use spatial-softmax temperature $Temp = 10$ when converting decoder logits to normalized gaze heatmaps and gaze centers. During reference construction, uncertainty feedback to UCTA is neutralized by using a temporally constant uncertainty profile. Because temporal standardization removes a constant profile, this recovers uncertainty-free cross-time attention and keeps the MC references independent of the current uncertainty estimates. The MC references are recomputed every four optimization steps, with the corresponding uncertainty loss rescaled to preserve its effective weight. Gradients are stopped through the MC references and estimator inputs.

Structured uncertainty objective. For compactness in this subsection, let $a \in \{\text{traj}, \text{hmap}\}$, and define

$$p_{nr}^{a} = \log \hat{u}_{nr}^{a}, \qquad t_{nr}^{a} = \log u_{nr}^{a,\text{MC}}, \tag{27}$$

for sample n and interval r. We omit the superscript a below for clarity. Let $\bar{p}_n = R^{-1} \sum_r p_{nr}$ and $\bar{t}_n = R^{-1} \sum_r t_{nr}$. The sample-level term is

$$\mathcal{L}_{\text{global}} = \text{SL1}(\bar{p}, \bar{t}), \tag{28}$$

where SL1 denotes Smooth-L_1 loss.

To preserve relative uncertainty across samples, let $\mathcal{P} = \{(i, j) : 1 \le i < j \le B\}$ and

$$s_g = \max\left(\text{Std}_{(i,j)\in\mathcal{P}}(\bar{t}_i - \bar{t}_j), s_{\min}\right). \tag{29}$$

We define

$$\mathcal{L}_{\text{global-pair}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \text{SL1}\left(\frac{\bar{p}_i - \bar{p}_j}{s_g}, \frac{\bar{t}_i - \bar{t}_j}{s_g}\right). \tag{30}$$

The shared interval-wise pattern is

$$p_r^{\text{temp}} = \frac{1}{B} \sum_n p_{nr} - \frac{1}{BR} \sum_{n,r} p_{nr}, \tag{31}$$

with t_r^{temp} defined analogously. Using $s_t = \max(\text{Std}(t), s_{\min})$,

$$\mathcal{L}_{\text{template}} = \frac{1}{R} \sum_r \text{SL1}\left(\frac{p_r^{\text{temp}}}{s_t}, \frac{t_r^{\text{temp}}}{s_t}\right). \tag{32}$$

Finally, to capture sample-specific temporal variation, define

$$p_{nr}^{\text{rel}} = p_{nr} - \bar{p}_n, \qquad t_{nr}^{\text{rel}} = t_{nr} - \bar{t}_n. \tag{33}$$

For each sample pair, we match differences between their relative interval profiles:

$$\mathcal{L}_{\text{pair}} = \frac{1}{|\mathcal{P}|R} \sum_{(i,j)\in\mathcal{P}} \sum_r \text{SL1}\left(\frac{p_{ir}^{\text{rel}} - p_{jr}^{\text{rel}}}{s_p}, \frac{t_{ir}^{\text{rel}} - t_{jr}^{\text{rel}}}{s_p}\right), \tag{34}$$

where s_p is the standard deviation of the corresponding target pair differences, lower-bounded by $s_{\min}$. We use $s_{\min} = 0.03$ and unit weights for the three additional structured terms. The uncertainty objective for form a is

$$\mathcal{L}_{\text{unc}}^{a} = \mathcal{L}_{\text{global}}^{a} + \mathcal{L}_{\text{global-pair}}^{a} + \mathcal{L}_{\text{template}}^{a} + \mathcal{L}_{\text{pair}}^{a}. \tag{35}$$

The final uncertainty loss is

$$\mathcal{L}_{\text{unc}} = \frac{1}{2}\left(\mathcal{L}_{\text{unc}}^{\text{traj}} + \mathcal{L}_{\text{unc}}^{\text{hmap}}\right). \tag{36}$$

A.4 Uncertainty-Aware Cross-Time Attention

For each spatial position p at decoder stage s, we collect the temporal feature sequence $\mathbf{H}_p^{(s)} \in \mathbb{R}^{T_s \times C_s}$ and apply pre-normalized multi-head self-attention. For head h,

$$\mathbf{Q}_{p,h}^{(s)} = \mathrm{LN}(\mathbf{H}_p^{(s)}) W_{Q,h}^{(s)}, \quad \mathbf{K}_{p,h}^{(s)} = \mathrm{LN}(\mathbf{H}_p^{(s)}) W_{K,h}^{(s)}, \quad \mathbf{V}_{p,h}^{(s)} = \mathrm{LN}(\mathbf{H}_p^{(s)}) W_{V,h}^{(s)}. \tag{37}$$

The content logits are

$$A_{p,h,\tau,\tau'}^{(s)} = \frac{\mathbf{Q}_{p,h,\tau}^{(s)} \mathbf{K}_{p,h,\tau'}^{(s)\top}}{\sqrt{d}}. \tag{38}$$

The final VANTAGE model uses trajectory-level uncertainty as the reliability signal:

$$\ell_r = \log \hat{u}_r^{\mathrm{traj}}. \tag{39}$$

At decoder stage s, each temporal position receives the score of its corresponding future interval:

$$\ell_\tau^{(s)} = \ell_{r_s(\tau)}. \tag{40}$$

We standardize this profile over time:

$$\tilde{\ell}_\tau^{(s)} = \frac{\ell_\tau^{(s)} - \mu_\ell^{(s)}}{\sigma_\ell^{(s)} + \epsilon}. \tag{41}$$

The standardized score is applied as a key-side attention bias:

$$\hat{A}_{p,h,\tau,\tau'}^{(s)} = A_{p,h,\tau,\tau'}^{(s)} - \tilde{\ell}_{\tau'}^{(s)}. \tag{42}$$

Thus, at fixed content logits,

$$\alpha_{p,h,\tau,\tau'}^{(s)} \propto \exp(A_{p,h,\tau,\tau'}^{(s)}) \exp(-\tilde{\ell}_{\tau'}^{(s)}), \tag{43}$$

so content determines temporal relevance while trajectory-level uncertainty modulates reliability.

The UCTA output is

$$\mathbf{O}_p^{(s)} = \mathrm{Proj}^{(s)} \left(\mathrm{Concat}_h \left[\mathrm{Softmax}(\hat{\mathbf{A}}_{p,h}^{(s)}) \mathbf{V}_{p,h}^{(s)} \right] \right). \tag{44}$$

It is integrated through a stage-specific gated residual connection:

$$\mathbf{H}^{(s)} \leftarrow \mathbf{H}^{(s)} + \gamma_s^{\mathrm{ucta}} \mathbf{O}^{(s)}, \tag{45}$$

where γ_s^{ucta} is a learnable stage-specific gate. UCTA is applied before interval-specific latent injection at every decoder stage. The uncertainty scores supplied to UCTA are detached from prediction-loss gradients, so the uncertainty estimator is trained by $\mathcal{L}_{\mathrm{unc}}$ rather than being redefined by the reconstruction objective. A temporally constant uncertainty profile has zero standardized bias and therefore reduces to uncertainty-free cross-time attention.

A.5 Learning Objective and Optimization

Posterior Best-of-K reconstruction. During training, we draw $K_{\mathrm{train}} = 4$ posterior samples and decode the corresponding future predictions, and see Appendix A.2

Latent regularization. For the continuous interval latents,

$$\mathcal{L}_{\mathrm{kl}}^z = \frac{1}{R} \sum_{r=1}^{R} D_{\mathrm{KL}} \left(q_\phi(z_r \mid c, G_f^{(r)}) \,\|\, p_\theta(z_r \mid c) \right). \tag{46}$$

The complete latent regularization is

$$\mathcal{L}_{\mathrm{kl}} = \mathcal{L}_{\mathrm{kl}}^z + \lambda_m \mathcal{L}_{\mathrm{kl}}^m + \lambda_b \mathcal{L}_{\mathrm{bal}}. \tag{47}$$

Diversity regularization. For two gaze trajectories decoded from different latent samples, we define their average pairwise trajectory distance as

$$\mathrm{Dist} = \frac{1}{t_f} \sum_{\tau} \left\| (x, y)_\tau - (x', y')_\tau \right\|_2. \tag{48}$$

The diversity regularization is

$$\mathcal{L}_{\text{div}} = \max\left(0,\, m_{\text{div}} - \text{Dist}\right), \tag{49}$$

where $(x, y)_\tau$ and $(x', y')_\tau$ are gaze centers decoded from two different latent samples.

Optimization. The final objective is

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda_{\text{KL}}\mathcal{L}_{\text{kl}} + \lambda_{\text{unc}}\mathcal{L}_{\text{unc}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}}. \tag{50}$$

All models are trained for 25 epochs with AdamW, batch size 8, and the learning-rate schedule reported in Sec. 4.1. The uncertainty references are evaluated every four optimization steps, as described in Appendix A.3.

B Evaluation Details

B.1 Predictive and Distributional Metrics

Let $(x, y)_\tau$ denote the ground-truth gaze coordinate at future time τ, and $(x, y)_{k,\tau}$ the soft-argmax coordinate recovered from stochastic prediction k.

ADE. For stochastic prediction k,

$$\text{ADE}_k = \frac{1}{|\Omega|}\sum_{\tau\in\Omega} \left\|(x, y)_{k,\tau} - (x, y)_\tau\right\|_2, \tag{51}$$

where Ω contains valid future fixation frames. We report

$$\text{minADE}\,@N = \min_{1\le k\le N} \text{ADE}_k, \qquad \text{meanADE}\,@N = \frac{1}{N}\sum_{k=1}^{N} \text{ADE}_k\,. \tag{52}$$

All probabilistic results use $N = 20$ stochastic predictions.

Negative log-likelihood. At each future time point, we fit a diagonal Gaussian to the N sampled gaze coordinates. For coordinate component $d \in \{x, y\}$, let $d_{k,\tau}$ and d_τ denote the corresponding predicted and ground-truth coordinates, respectively. We compute

$$\bar{d}_\tau = \frac{1}{N}\sum_{k=1}^{N} d_{k,\tau}, \qquad s_{\tau,d} = \text{Std}_k(d_{k,\tau}). \tag{53}$$

The per-time-point NLL is

$$\text{NLL}_\tau = \frac{1}{2}\sum_{d\in\{x,y\}} \left[\left(\frac{d_\tau - \bar{d}_\tau}{s_{\tau,d} + \epsilon}\right)^2 + 2\log(s_{\tau,d} + \epsilon) + \log(2\pi)\right]. \tag{54}$$

NLL is averaged over valid future fixation frames and reported in pixel space.

CRPS. We use the empirical coordinate-wise CRPS. For $d \in \{x, y\}$,

$$\text{CRPS}_{\tau,d} = \frac{1}{N}\sum_{k=1}^{N} |d_{k,\tau} - d_\tau| - \frac{1}{2N^2}\sum_{k=1}^{N}\sum_{\ell=1}^{N} |d_{k,\tau} - d_{\ell,\tau}|. \tag{55}$$

We sum the two coordinate scores and average over valid future fixation frames. Lower NLL and CRPS indicate better distributional quality.

B.2 Uncertainty Fidelity Metrics

We evaluate each uncertainty estimate against its corresponding $K = 20$ output-space MC reference using Spearman rank correlation.

Overall correlation. Overall ρ is computed over all sample–interval pairs:

$$\rho_{\text{overall}} = \rho_{\text{S}}\left(\{\hat{u}_{nr}\}_{n,r}, \{u^{\text{MC}}_{nr}\}_{n,r}\right). \tag{56}$$

Sample correlation. For each sample, we first average uncertainty over future intervals:

$$\bar{u}_n = \frac{1}{R} \sum_r u_{nr}. \tag{57}$$

Sample ρ is the Spearman correlation between $\{\bar{\hat{u}}_n\}_n$ and $\{\bar{u}_n^{\mathrm{MC}}\}_n$.

Temporal correlation. To isolate sample-specific temporal variation, we remove sample-level and shared interval effects. For a matrix $U = [u_{nr}]$, define the double-centered residual

$$U_{nr}^{\mathrm{temp}} = u_{nr} - \bar{u}_{n\cdot} - \bar{u}_{\cdot r} + \bar{u}_{\cdot\cdot}. \tag{58}$$

Temporal ρ is the Spearman correlation between the flattened residual matrices of the estimated and MC uncertainties. This measures whether the estimator recovers how uncertainty varies across intervals for individual samples rather than only a shared horizon-dependent profile.

For Raw-σ, the interval score is the mean conditional-prior scale over latent dimensions:

$$u_{nr}^{\mathrm{raw}} = \frac{1}{d_z} \sum_{j=1}^{d_z} \sigma_{nr,j}^{p}. \tag{59}$$

B.3 Temperature-Sensitivity Diagnostic

The temperature analysis isolates the semantic role of heatmap-level uncertainty. We use the same final Pred-Traj checkpoint and the same fixed realized interval ADEs as in Appendix C.1. For each batch, raw stochastic decoder outputs are sampled once and reused at all spatial-softmax temperatures

$$Temp \in \{5, 7.5, 10, 12.5, 20\}. \tag{60}$$

Changing $Temp$ therefore changes only the spatial normalization applied to the fixed decoder outputs; latent variables, mode variables, and realized errors are not resampled.

For a spatially normalized predicted gaze heatmap $\hat{g}_{k,\tau}$, we additionally compute normalized spatial entropy

$$H_{\mathrm{norm}}(\hat{g}_{k,\tau}) = -\frac{\sum_{i,j} \hat{g}_{k,\tau}(i,j) \log \hat{g}_{k,\tau}(i,j)}{\log(HW)}. \tag{61}$$

Entropy is averaged over stochastic samples and future time points within each interval. We report its rank correlation with MC heatmap-level uncertainty to characterize whether heatmap-level uncertainty follows within-heatmap spatial dispersion.

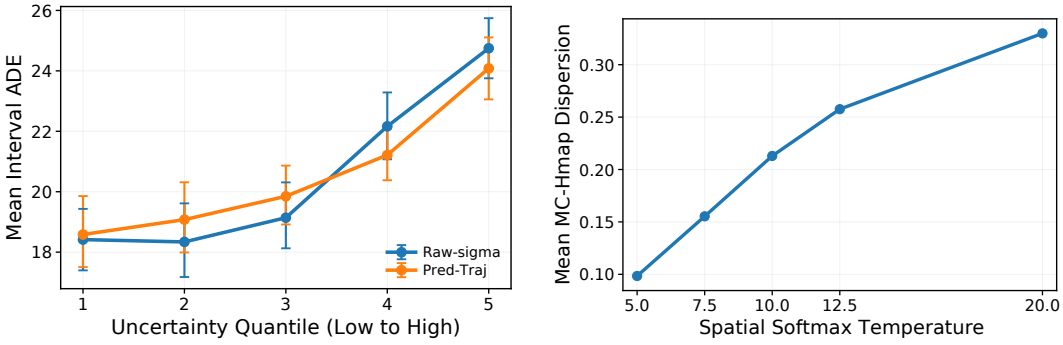
C Additional Uncertainty Analysis

This section provides additional diagnostics for the reliability and semantic interpretation of the predictive uncertainty analyzed in the main text. We examine whether trajectory-level uncertainty remains associated with realized forecasting error after controlling for forecast interval, and how heatmap-level uncertainty responds to changes in spatial decoding.

C.1 Interval-Stratified Reliability and Heatmap Dispersion

Figure 4(a) further examines the realized-error association of trajectory-level uncertainty. Because forecasting error generally varies over the forecast horizon, we form uncertainty quantiles separately within each interval before aggregating samples with the same quantile index. The mean realized ADE increases monotonically across the five Pred-Traj uncertainty quantiles ($18.58, 19.07, 19.86, 21.20$, and 24.08). This trend indicates that trajectory-level uncertainty distinguishes relatively reliable and unreliable predictions beyond differences in baseline difficulty across forecast intervals, complementing the risk–coverage analysis in the main text.

Figure 4(b) examines the magnitude of heatmap-level output-space MC uncertainty under different spatial-softmax temperatures. Increasing the temperature from $Temp = 5$ to $Temp = 20$ increases the mean heatmap-level MC uncertainty from 0.098 to 0.330. This sensitivity is consistent with its definition as within-heatmap spatial dispersion, which depends directly on the spatial normalization applied to the decoder outputs.



(a) Interval-stratified reliability

(b) Heatmap-level dispersion

Figure 4: Additional analysis of predictive uncertainty on Ego-Exo4D. (a) Mean realized interval ADE across uncertainty quantiles formed separately within each forecast interval. (b) Mean heatmap-level output-space MC uncertainty $u^{\text{hmap,MC}}$ under varying spatial-softmax temperature. Error bars denote 95% take-level bootstrap confidence intervals where applicable.

Table 6: Temperature-sensitivity analysis on Ego-Exo4D. All quantities are computed from the same checkpoint, fixed stochastic decoder outputs, and fixed realized interval errors.

$Temp$	$\rho(u^{\text{hmap,MC}}, \text{ADE})$	$\text{AURC}(u^{\text{hmap,MC}})$	$\rho(u^{\text{traj,MC}}, \text{ADE})$	$\rho(u^{\text{hmap,MC}}, H_{\text{norm}})$
5.0	0.097	19.60	0.412	0.838
7.5	-0.010	20.78	0.386	0.798
10.0	-0.100	21.65	0.314	0.782
12.5	-0.167	22.27	0.258	0.758
20.0	-0.276	23.33	0.199	0.675

C.2 Temperature-Sensitivity Results

Table 6 provides a more detailed analysis of the semantic roles of trajectory- and heatmap-level uncertainty under changes in spatial-softmax temperature. All quantities are computed from the same checkpoint, fixed stochastic decoder outputs, and fixed realized interval errors, such that changing $Temp$ affects only the spatial normalization of the predicted heatmaps.

As $Temp$ increases from 5 to 20, the correlation between heatmap-level MC uncertainty and realized ADE changes from 0.097 to -0.276, while its AURC increases from 19.60 to 23.33. In contrast, trajectory-level MC uncertainty remains positively associated with realized ADE throughout the temperature sweep ($\rho = 0.199$–0.412).

Heatmap-level uncertainty exhibits a different semantic behavior. Its correlation with normalized spatial entropy remains consistently high ($\rho = 0.675$–0.838), indicating that it primarily characterizes within-heatmap spatial dispersion. The fidelity analysis in Table 5 separately shows that the learned heatmap-level uncertainty estimate remains aligned with its corresponding output-space MC reference. Therefore, the weak and temperature-sensitive association between heatmap-level dispersion and realized error is not explained simply by poor estimator fidelity. Thus, the weaker association of heatmap-level uncertainty with realized error cannot be attributed simply to failure of the uncertainty estimator.

Together, these diagnostics reinforce the distinction observed in the main text. Trajectory-level uncertainty is more consistently associated with realized forecast reliability, whereas heatmap-level uncertainty characterizes within-heatmap spatial spread and is sensitive to the spatial decoding scale. This difference supports treating the two uncertainty forms according to their semantic roles rather than collapsing them into a single reliability signal.