SPECTRA-GAZE: GAZE TARGET ESTIMATION VIA HEAD-CONDITIONED SPECTRAL MODULATION AND RELIABILITY-AWARE HEATMAP REFINEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Gaze target estimation combines head cues and scene information to determine where a person is looking. When a scene contains several plausible targets, the same head pose may not distinguish between them, and distractors may receive strong heatmap responses. Existing methods mainly improve head–scene interaction in the spatial domain, but do not explicitly address ambiguity both before and after heatmap decoding. We view gaze ambiguity as two linked problems: weak target evidence before decoding and unreliable candidate peaks afterward. We propose SPECTRA-Gaze to address both through Head-Conditioned Spectral State Interaction (HSSI) and Reliability-Constrained Nonlocal State Interaction (RCNSI). When nearby plausible targets are difficult to distinguish from head pose alone, scene cues around them become important. The DCT separately represents broad and fine spatial variation in scene features. HSSI forms adaptive frequency bands from scene and head descriptors, allowing head cues to modulate scene details that may help distinguish nearby plausible targets. Interactions among band descriptors then guide spectral processing before decoding. RCNSI uses candidate reliability to weight candidate features alongside pooled scene context and refine the coarse heatmap. A spatial gate combines local heatmap dispersion with reliable candidate support to control the location and magnitude of logit corrections. Both modules use a partially shared dual-stream encoder to reduce parameter redundancy. Extensive experiments on the GazeFollow and VideoAttentionTarget datasets demonstrate that our method consistently outperforms state-of-the-art methods while using 18.05% fewer parameters.

Keywords: Gaze target estimation; Spectral modulation; Nonlocal reasoning; Candidate reliability.

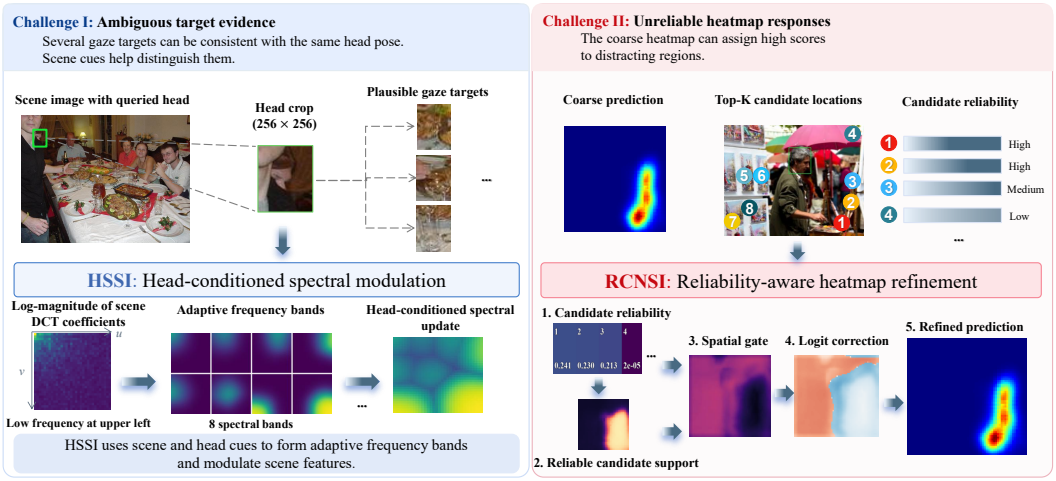

Figure 1: Motivation for SPECTRA-Gaze. HSSI addresses ambiguous target evidence through head-conditioned spectral modulation. RCNSI uses candidate reliability to guide refinement of the coarse heatmap.

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.

1 INTRODUCTION

Knowing where a person is looking helps us understand interactions between people and between people and objects. Gaze target estimation predicts where a specified person is looking from a scene image and the person's head bounding box (Recasens et al., 2015; Chong et al., 2020). This task requires interpreting head cues in the context of the scene. Prior studies explored this relationship through joint gaze–scene attention (Chong et al., 2018; Fang et al., 2021), scene geometry (Bao et al., 2022; Jin et al., 2022), and gaze-distribution modeling (Miao et al., 2023; Lian et al., 2019). However, similar head poses can correspond to several plausible targets. Resolving this gaze ambiguity requires identifying target-relevant scene cues and separating them from distracting information.

Recent research highlights the importance of how head cues enter a gaze target estimation model. Gaze-LLE (Ryan et al., 2025) decodes gaze targets from shared pretrained scene features using a person-location prompt. Concurrent PaGE (Ye et al., 2026) introduces explicit interactions between scene features and head features. These methods improve head–scene interaction, but they do not clearly distinguish between two stages: how to form a scene representation before decoding, and how to evaluate competing heatmap candidates after decoding. Before decoding, weak target cues may be masked by other scene information. After decoding, peaks with strong responses may also point to distracting locations. This raises the question: How can we strengthen target cues before decoding and use candidate reliability to refine the heatmap after decoding?

We view gaze ambiguity through two interrelated considerations: which scene cues should be emphasized before decoding, and how much weight should be assigned to information at each candidate location during heatmap refinement. We therefore use head cues to modulate scene features across frequency bands. The discrete cosine transform (DCT) represents these features as frequency components that can be adjusted separately (Qin et al., 2021). Adaptive frequency bands use joint scene and head descriptors to determine which components are processed together. This allows the queried person's head cues to guide scene-feature modulation before heatmap decoding.

SPECTRA-Gaze handles these two decisions via HSSI and RCNSI respectively. Head-Conditioned Spectral State Interaction (HSSI) modulates scene frequencies with head information to strengthen weak target cues before decoding. Reliability-Constrained Nonlocal State Interaction (RCNSI) weights candidate features by reliability during heatmap refinement, limiting the influence of distracting locations. Pooled scene context provides additional information, while a spatial gate controls the resulting logit corrections. Both modules use a compact, partially shared dual-stream encoder based on DINOv3 (Siméoni et al., 2025).

We evaluate SPECTRA-Gaze on GazeFollow and VideoAttentionTarget against Gaze-LLE and concurrent PaGE. Relative to the published Gaze-LLE ViT-L results, SPECTRA-Gaze reduces average GazeFollow L_2 by 14.1% and VideoAttentionTarget L_2 by 23.7%. Under the same training and evaluation protocol, our method outperforms PaGE ViT-B without distillation across all six metrics, with 18.05% fewer parameters. Component ablations on the same compact base network examine the respective contributions of HSSI and RCNSI.

Our main contributions are: **(a)** We formulate gaze ambiguity as a two-stage problem: target evidence may be weak before decoding, and candidate context may be unreliable during heatmap refinement. **(b)** We propose HSSI, which uses head cues to construct adaptive frequency bands and guide their interactions, conditioning spectral processing on the queried person. **(c)** We propose RCNSI, which uses learned candidate reliability to guide context aggregation and a spatial gate to control logit corrections.

2 RELATED WORK

Gaze target estimation. Early gaze-target models fused head and scene features (Recasens et al., 2015; Chong et al., 2018; 2020). Later work added dual attention (Fang et al., 2021) and 3D scene geometry (Bao et al., 2022). Distribution-based methods represent multiple possible gaze targets (Miao et al., 2023; Li et al., 2026); other studies introduce object, temporal, or social-interaction cues (Tonini et al., 2023; Recasens et al., 2017; Gupta et al., 2024a). Other work studies broader settings, including domain shifts, children's gaze, conversational audio, multi-person temporal scenes, and multi-view inputs (Tonini et al., 2022; Tafasca et al., 2023; Hou et al., 2024; Gupta et al.,

2024b; Miao et al., 2025). Other methods extend gaze prediction toward semantic, pixel-level, and object-aware targets (Tafasca et al., 2024a; Liu et al., 2024a; Nieva-Suárez et al., 2025). Textual scene cues and language reasoning provide another source of context (She et al., 2026; Yang & Lu, 2025), while semi-supervised annotations and concept-based queries expand the training and input settings (Miao et al., 2024; Cao et al., 2026). Gaze-LLE (Ryan et al., 2025) first extracts pretrained DINOv2 scene features (Oquab et al., 2024), then introduces the queried person's information through a person-location prompt.PaGE (Ye et al., 2026) uses a separate head stream and explicit scene–head interaction layers on DINOv3 features (Siméoni et al., 2025).

SPECTRA-Gaze retains the person-location prompt used by Gaze-LLE and the explicit scene–head interaction layers used by PaGE. Unlike PaGE, its scene and head streams share some encoder parameters. Furthermore, it modulates scene frequency components based on the queried person's head cues before interaction and weights coarse-heatmap candidate features by learned reliability after decoding.

Conditional and spectral feature modulation. FiLM (Perez et al., 2018) showed that conditioning inputs can modulate feature responses through feature-wise affine transformations. Spectral methods modulated visual features through DCT-based channel attention (Qin et al., 2021) and fusion of local and spectral features (Chi et al., 2020). Related spectral models use global filtering (Rao et al., 2021), adaptive mixing (Guibas et al., 2022), or input-dependent filters (Tatsunami & Taki, 2024).

HSSI combines conditional modulation and spectral processing for the queried person: scene and head descriptors jointly determine soft DCT bands. Interactions among band descriptors then guide spectral processing before scene–head interaction.

State-space processing and context selection. Structured state-space models (Gu et al., 2022) and input-dependent selective updates (Gu & Dao, 2024) provided a foundation for modeling long-range dependencies. Vision models extended these mechanisms through bidirectional processing (Zhu et al., 2024) and two-dimensional selective scanning (SS2D) (Liu et al., 2024b). Nonlocal networks (Wang et al., 2018) aggregated spatial information using pairwise feature affinities.

RCNSI draws on both ideas for coarse-heatmap refinement, but further evaluates the reliability of each candidate source. Target-proximity supervision trains a reliability predictor, whose output weights candidate features. The zero-valued null route allows the model to withhold context, while the spatial gate restricts where that context can modify the heatmap logits.

3 METHOD

3.1 PROBLEM FORMULATION AND OVERVIEW

Given a scene image I and a person's head bounding box b, gaze target estimation predicts a heatmap $H \in [0,1]^{64\times64}$ and an in-frame probability p_{in}. Gaze ambiguity arises at two points in the pipeline. Before decoding, target cues may remain weak in scene features; after decoding, the strongest response in the coarse heatmap may belong to a distractor. HSSI addresses the first issue through head-conditioned spectral modulation. RCNSI addresses the second issue by using candidate reliability to guide context aggregation and refine the coarse heatmap. Figure 2 shows the overall framework.

We use a compact dual-stream encoder based on DINOv3 ViT-B/16 (Siméoni et al., 2025). The scene and head streams share input embeddings, positional encodings, and the first four encoder blocks. The remaining blocks and projection layers are separate for each stream and produce 256-channel scene features S and head features Q. HSSI modulates S using information from both streams. A person-location prompt and three bidirectional interaction layers then exchange scene and head information. Each stream includes four register tokens and one in/out token. The scene decoder produces coarse logits L, while the two in/out tokens jointly predict p_{in}. RCNSI refines only L and leaves the in/out prediction unchanged.

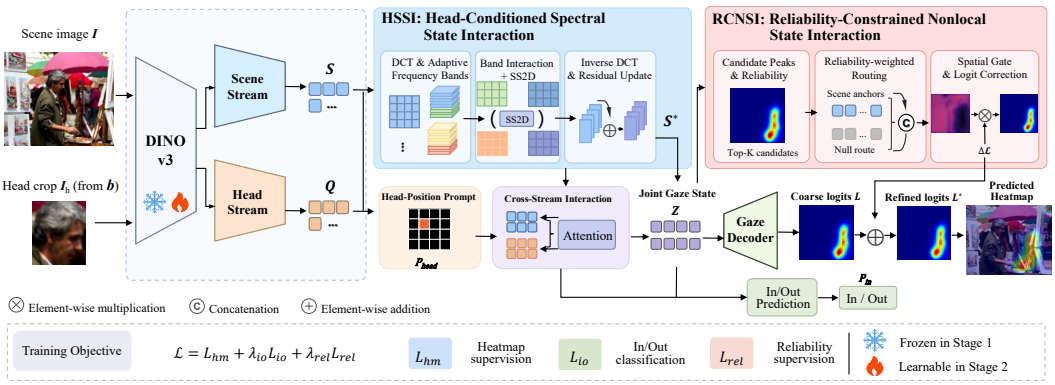

Figure 2: Overview of SPECTRA-Gaze. Head-Conditioned Spectral State Interaction (HSSI) modulates scene features before cross-stream interaction; Reliability-Constrained Nonlocal State Interaction (RCNSI) refines the coarse heatmap using candidate reliability, pooled scene context, and a spatial gate.

3.2 Head-Conditioned Spectral State Interaction

Before decoding, competing scene information may obscure weak target cues. Both broad scene structure and fine local details may help distinguish several plausible targets. The DCT represents these variations separately, allowing the queried person's head cues to guide how they are processed. HSSI uses scene and head descriptors to construct adaptive frequency bands. Interactions among band descriptors combine evidence from different bands. Two-dimensional selective scanning (SS2D) (Liu et al., 2024b) processes dependencies across the frequency grid. Finally, the inverse DCT maps the spectral update back to the spatial domain and adds it to the scene features as a residual.

Head-conditioned frequency bands. The frequency components relevant to a target can vary with the scene and the queried person. We therefore use scene and head descriptors to predict soft band masks. Let $\mathcal{D}$ and $\mathcal{D}^{-1}$ denote the orthonormal two-dimensional DCT and its inverse. We project S and Q to 96 channels and transform them into spectra $\widehat{F}_s$ and $\widehat{F}_h$. We globally average their log-magnitudes and concatenate the resulting descriptors. An MLP maps the concatenated descriptors to parameters $P \in \mathbb{R}^{K\times 5}$ for $K = 8$ bands. Band $k = 1, \ldots, K$ has center μ_k, scales d_k, and shear o_k. In $P_{k,1:2}$, $P_{k,3:4}$, and $P_{k,5}$, the subscripts select row k and the indicated entries. The constants α_c, $d_{\min}$, $d_{\max}$, and $o_{\max}$ bound center shifts, scales, and shear. We use σ for the sigmoid and $\tanh$ for the hyperbolic tangent. The reference-center logit is $\widetilde{\mu}_k^0$. The bounded parameters are

$$\begin{cases} \mu_k = \sigma\Big[\widetilde{\mu}_k^0 + \alpha_c \tanh(P_{k,1:2})\Big], \\ d_k = d_{\min} + (d_{\max} - d_{\min})\,\sigma(P_{k,3:4}), \\ o_k = o_{\max} \tanh(P_{k,5}). \end{cases} \tag{1}$$

The bounded center, scale, and shear parameters define anisotropic Gaussian responses over normalized frequency coordinates. At each frequency coordinate u, we normalize the responses across bands to obtain soft masks $M_k(u)$ that sum to one. These masks are then used to extract band descriptors and modulate the scene spectrum. Subsequent filtering updates only the scene spectrum. Appendix B gives the mask construction, parameter bounds, and reference centers.

Interactions among frequency bands. Within each band, we compute a mask-weighted mean of the scene log-magnitudes. We add a learned embedding of the band's center, scales, and shear to form a 96-dimensional band descriptor. Layer normalization and four-head self-attention exchange information among the descriptors through a residual connection. Let $\widetilde{z}_k$ be the updated band descriptor and f_ν an affine map for stage $\nu \in \{\mathrm{pre}, \mathrm{post}\}$. Their channel-wise modulation at frequency

u is $P_\nu(u)$:

$$P_\nu(u) = \sum_{k=1}^{K} M_k(u)\ \tanh[f_\nu(\widetilde{z}_k)]\,, \qquad \nu \in \{\mathrm{pre}, \mathrm{post}\}. \tag{2}$$

Each frequency position combines modulation from multiple bands using the soft masks. The resulting P_{pre} and P_{post} act before and after frequency-domain scanning, respectively.

Spectral processing and residual update. Let $\mathcal{S}_f$ denote an SS2D adapter with input group normalization and four directional scans. The sign-preserving transform $\psi(x) = \mathrm{sign}(x)\log(1+|x|)$ compresses the scan input. The modulation strengths are $\lambda_{\mathrm{pre}} = \lambda_{\mathrm{post}} = 0.25$, and $\odot$ denotes element-wise multiplication. The spectral gate G is

$$G = \frac{1}{2}\tanh\Big\{\mathcal{S}_f\Big[\psi(\widehat{F}_s) \odot \big(1 + \lambda_{\mathrm{pre}} P_{\mathrm{pre}}\big)\Big]\Big\}. \tag{3}$$

Thus, SS2D processes the compressed spectrum, while $\tanh$ bounds the spectral gate. The 1×1 projection W_{write} maps the inverse-DCT result back to scene features. The scalar residual weight is $g_s = \tanh(\gamma)$, with γ initialized to zero. The updated scene feature $S^\star$ is

$$S^\star = S + g_s W_{\mathrm{write}}\, \mathcal{D}^{-1}\Big[\widehat{F}_s \odot G \odot \big(1 + \lambda_{\mathrm{post}} P_{\mathrm{post}}\big)\Big]. \tag{4}$$

The residual weight is zero at initialization, so HSSI initially preserves the scene features. The updated $S^\star$ then enters the scene–head interaction layers before heatmap decoding.

3.3 Reliability-Constrained Nonlocal State Interaction

A high-response location in the coarse heatmap may be the gaze target or a distractor. RCNSI therefore separates candidate proposal from candidate weighting. The coarse heatmap proposes candidate locations, while learned candidate reliability weights their features during heatmap refinement. Candidate peaks retain local target hypotheses, while pooled scene anchors provide context beyond those peaks. A null route and a spatial gate control how the aggregated information modifies the logits.

Candidate and scene memory. We average-pool the coarse logits L to 32×32 and apply a spatial softmax to obtain an auxiliary distribution p. This distribution is used for candidate selection; the final heatmap applies a sigmoid independently to the refined logits at each location. Let $x_i \in \mathbb{R}^{96}$ denote the aligned scene feature at receiver grid cell i, and let $h \in \mathbb{R}^{96}$ be the pooled head feature. We prioritize 3×3 local maxima in p and select the eight positions with the highest masked scores. For each selected candidate $k \in \{1, \ldots, 8\}$, an MLP predicts a logit ρ_k. Its sigmoid $a_k = \sigma(\rho_k)$ is the candidate reliability.

The MLP receives the candidate feature, h, head-box displacement, local probability mass, normalized entropy, and peak margin. Candidate reliability estimates graded proximity to the gaze target and supplies a prior for weighting candidate sources during context aggregation. Appendix E examines its relationship with candidate localization quality.

We combine the eight candidates with 16 scene anchors from 4×4 regional pooling to form memory $\{(v_j, c_j)\}_{j=1}^{24}$. Each entry contains a feature v_j and normalized coordinate c_j; $j = 1, \ldots, 8$ indexes candidates and $j = 9, \ldots, 24$ indexes scene anchors.

Reliability-weighted routing with a null source. We assign each candidate its reliability $s_j = a_j$ and each scene anchor a fixed prior $s_j = 0.25$. We project $x_i + h$ and v_j into ℓ_2-normalized queries q_i and keys k_j. For receiver i and source j, the routing logit ℓ_{ij} combines feature similarity, relative geometry, and source reliability. The receiver coordinate is ξ_i, and ϕ is a learned geometric bias. We use similarity scale $\kappa = 4$ and prior clamp $\epsilon = 10^{-8}$:

$$\ell_{ij} = \underbrace{\kappa\, q_i^{\mathsf{T}} k_j}_{\text{feature similarity}} + \underbrace{\phi(\xi_i - c_j)}_{\text{relative geometry}} + \underbrace{\log\big(\max\{s_j, \epsilon\}\big)}_{\text{source reliability}}. \tag{5}$$

The log prior downweights low-reliability candidates, while feature similarity and relative geometry measure their relevance to each receiver. We add a null source with a zero feature vector and a

learnable logit b_{null}, initialized to 2. A softmax over the 24 real-source logits and the null logit produces routing weights r_{ij} and $r_{i,\text{null}}$. With learned value projection W_v, the message $m_i = \sum_{j=1}^{24} r_{ij} W_v v_j$ aggregates the real-source features. Since the null source contributes a zero feature vector, assigning more weight to it attenuates the aggregated context. We add m_i to x_i and apply a residual spatial SS2D adapter. Decoding and bilinear upsampling produce a correction map $R \in \mathbb{R}^{64\times64}$.

Spatially constrained logit refinement. The aggregated context should not modify every location equally. The spatial gate combines local heatmap dispersion and reliable candidate support to control where and how strongly the context changes the logits. We renormalize the probabilities within each 3×3 neighborhood and divide their entropy by $\log 9$ to obtain local dispersion e_i. Larger e_i means the local probability mass is more dispersed. Reliable candidate support is $u_i = \sum_{k=1}^{8} r_{ik} a_k$. A second MLP predicts a sigmoid factor τ_i from the receiver features, e_i, u_i, null-route weight $r_{i,\text{null}}$, and auxiliary probability p_i at receiver i. We also stop gradients through these inputs. The gate is $T_i = \tau_i e_i \sqrt{u_i}$, and T collects its values on the receiver grid. Here $\overline{R}$ is the spatial mean of R, $\mathcal{U}$ bilinear upsampling, and $\beta = 0.5$ the correction scale. With η initialized to zero, $g_r = \tanh(\eta)$ weights the correction to form logits $L^\star$:

$$L^\star = L + \beta g_r \Big[\tanh\big(R - \overline{R}\big) \odot \mathcal{U}(T)\Big]. \tag{6}$$

Subtracting the spatial mean removes a global offset, $\tanh$ bounds the residual, and the spatial gate controls where it applies. Since $g_r = 0$ at initialization, the refined logits initially equal the coarse logits. The refined heatmap is $H = \sigma(L^\star)$.

3.4 Training objective

The heatmap and in/out losses supervise the two task outputs. RCNSI learns candidate reliability from soft target-proximity labels to weight candidate context. Let $\mathcal{L}_{\text{hm}}$ be pixelwise binary cross-entropy on heatmap logits for in-frame samples. Let $\mathcal{L}_{\text{io}}$ be binary cross-entropy on the in/out logit for all samples. We measure the distance between each detached candidate coordinate and the target-heatmap peak. A unit-peak Gaussian with standard deviation 0.08 gives the soft label. Coordinates are normalized to $[0, 1]^2$, with the target peak index divided by 63. Binary cross-entropy between the reliability logits and these soft labels defines $\mathcal{L}_{\text{rel}}$ on in-frame samples. We weight the in/out and reliability losses by λ_{io} and λ_{rel}:

$$\mathcal{L} = \mathcal{L}_{\text{hm}} + \lambda_{\text{io}}\mathcal{L}_{\text{io}} + \lambda_{\text{rel}}\mathcal{L}_{\text{rel}}. \tag{7}$$

We set $\lambda_{\text{io}} = \lambda_{\text{rel}} = 0.01$ and compute all binary cross-entropy terms directly from logits. Ablations without RCNSI omit $\mathcal{L}_{\text{rel}}$; HSSI requires no auxiliary supervision. We first train task-specific modules for 15 epochs with the backbone frozen, then jointly fine-tune for five epochs. Appendix A gives the remaining training details.

4 Experiments

4.1 Datasets and protocol

We evaluate SPECTRA-Gaze on GazeFollow (GF) and VideoAttentionTarget (VAT). GF pairs still images and queried head boxes with annotated gaze targets (Recasens et al., 2015). VAT provides annotated video frames and includes in-frame and out-of-frame gaze targets (Chong et al., 2020). GF reports heatmap AUC, average L_2 distance to the mean annotation, and minimum L_2 distance to any annotation. VAT reports AUC and L_2 on in-frame samples and in/out AP on all samples. All locally trained configurations shared DINOv3 pretrained weights, training data, data augmentation, the two-stage optimization schedule, and evaluation code. Before training, we fixed Stage 2 epoch 4 as the evaluation checkpoint for every configuration. Predicted coordinates were obtained by dividing heatmap peak indices by the heatmap resolution. Each configuration was jointly trained on the combined GF and VAT training sets, then evaluated on both test sets using the same checkpoint. Training used neither ChildPlay data nor unlabeled distillation data. Appendix A details the hardware, preprocessing, optimization, and coordinate conventions.

Table 1: Comparison on GazeFollow and VideoAttentionTarget. Published baseline results are displayed to three decimal places; the matched local results use four decimal places. Teal rows identify the matched local comparison, with PaGE denoting our ViT-B reproduction. **Bold red** marks the best and underline the second-best distinct model value per column; tied values share a rank. Higher AUC/AP and lower L_2 are better.

	GazeFollow			VideoAttentionTarget		
Method	AUC ↑	Avg L_2 ↓	Min L_2 ↓	AUC ↑	L_2 ↓	AP ↑
Chong et al. (2020)	0.921	0.137	0.077	0.860	0.134	0.853
Fang et al. (2021)	0.922	0.124	0.067	0.905	0.108	0.896
ESCNet (Bao et al., 2022)	0.928	0.122	–	0.885	0.120	0.869
Jin et al. (2022)	0.920	0.118	0.063	0.900	0.104	0.895
Gupta et al. (2022)	0.943	0.114	0.056	0.914	0.110	0.879
PDP (Miao et al., 2023)	0.934	0.123	0.065	0.917	0.109	0.908
Sharingan (Tafasca et al., 2024b)	0.944	0.113	0.057	–	0.107	0.891
ViTGaze (Song et al., 2024)	0.949	0.105	0.047	0.938	0.102	0.905
Gaze-LLE ViT-B (Ryan et al., 2025)	0.956	0.104	0.045	0.933	0.107	0.897
Gaze-LLE ViT-L (Ryan et al., 2025)	0.958	0.099	0.041	0.937	0.103	0.903
GazeVLM (Mathew et al., 2025)	0.929	0.131	0.076	0.926	0.112	0.898
Mi et al. (2026)	0.961	0.094	0.038	0.948	0.095	0.923
PaGE ViT-B (Ye et al., 2026)	0.9626	0.0887	0.0338	0.9595	0.0807	0.9378
SPECTRA-Gaze (ours)	**0.9630**	**0.0851**	**0.0317**	**0.9621**	**0.0785**	**0.9389**

4.2 Comparison with prior methods

Table 1 compares SPECTRA-Gaze with Gaze-LLE, concurrent PaGE, and earlier gaze target estimation methods. Published results retain their original training and decoding protocols; the two teal rows use our shared training and evaluation protocol. Relative to the published Gaze-LLE ViT-L results, SPECTRA-Gaze reduces average GF L_2 from 0.0990 to 0.0851 and VAT L_2 from 0.1030 to 0.0785. These correspond to relative error reductions of 14.1% and 23.7%. We next compare SPECTRA-Gaze and PaGE ViT-B under the same protocol.

Matched comparison. Under our shared implementation and evaluation protocol, PaGE ViT-B contains 187.306M parameters, compared with 153.501M in SPECTRA-Gaze, a reduction of 18.05%. On GF, average and minimum L_2 decreased by 0.0036 and 0.0021, respectively, while AUC increased by 0.0004. On VAT, L_2 decreased by 0.0022, while AUC and in/out AP increased by 0.0026 and 0.0011, respectively. Table 1 provides the complete results. Appendix D.3 reports means and standard deviations across three training seeds for both models. On GF, paired bootstrap intervals for both localization-error differences lie below zero; Appendix D.2 gives the paired comparison.

4.3 Qualitative analysis

Figure 3 compares Gaze-LLE, PaGE, and SPECTRA-Gaze on the same six examples. Across these examples, Gaze-LLE and PaGE often spread their responses over several plausible regions. Their peaks sometimes fall away from the annotated target. SPECTRA-Gaze produces a more compact response and places its peak closer to the target. In the VideoAttentionTarget sample on the far right, the predicted peaks of Gaze-LLE and PaGE fall on different people, while SPECTRA-Gaze's predicted peak is closer to the annotated gaze target.

4.4 Component ablation

Table 2 examines the compact base network and the two added modules separately. Base uses a partially shared dual-stream encoder and three interaction layers, with 153.094M parameters compared with 187.306M in PaGE ViT-B. HSSI and RCNSI add 0.206M and 0.201M parameters, respectively,



Figure 3: Qualitative comparison on GazeFollow and VideoAttentionTarget. Rows show Gaze-LLE ViT-L, PaGE ViT-B, SPECTRA-Gaze, and GT. Each column uses the same image and queried person. Green boxes mark the queried head; green points show gaze annotations in the GT row and predicted heatmap peaks in the remaining rows. Lines connect head centers to predicted peaks. The three predicted heatmaps in each column share a color scale.

bringing the full model to 153.501M. Base, HSSI-only, RCNSI-only, and the full model were trained independently.

Both single-module configurations improved all five localization metrics over Base. HSSI-only achieved lower average GF L_2 than RCNSI-only, while RCNSI-only achieved lower VAT L_2. Combining both modules gave the lowest localization errors on both datasets. Average GF L_2 decreased from 0.0890 for Base to 0.0851 for the full model, while VAT L_2 decreased from 0.0827 to 0.0785. VAT in/out AP decreased from 0.9439 to 0.9389; Base achieved the highest AP among these configurations. Configurations with RCNSI also used the auxiliary reliability loss defined in Section 3.4.

Figure 4 compares PaGE, w/o RCNSI, w/o HSSI, and SPECTRA-Gaze against the GT annotations. In these examples, PaGE often spreads its response over competing regions, and its peak can miss the annotated target. Without RCNSI, some secondary peaks remain; without HSSI, responses can spread across nearby targets. With both modules, SPECTRA-Gaze produces a more concentrated response and places its peak closer to the annotation in scenes with nearby people and objects. Appendix A.5 provides visualization details and internal-state analysis.

Candidate reliability. On both datasets, higher candidate reliability corresponded to higher candidate hit rates across reliability intervals (Figure 6; Appendix E). We also compared candidate reliability and coarse-heatmap scores on identical candidate pools with $K \in \{8, 10, 100\}$. Selecting the top-ranked candidate by reliability gave a lower average target distance on both datasets (Table 10). Appendix E.1 gives the full comparison. Figure 7 shows examples where candidate reliability selects a location closer to the annotated target than the competing heatmap peak.

5 Discussion

The ablations support complementary roles for head-conditioned spectral modulation and reliability-aware heatmap refinement. HSSI-only achieved lower average L_2 on GazeFollow than RCNSI-only, while RCNSI-only achieved lower L_2 on VAT. Combining the two gave the lowest localization errors on both datasets. These results suggest that gaze localization benefits from improvements to both scene features and decoded predictions.

The fixed-weight HSSI interventions in Appendix B.4 show higher localization error when the queried person's head descriptor is mismatched or adaptive frequency processing is altered. Using the same full-model checkpoint, RCNSI further reduces localization error after the coarse heatmap is generated (Appendix D.1). It modulates the contribution of candidate information to context

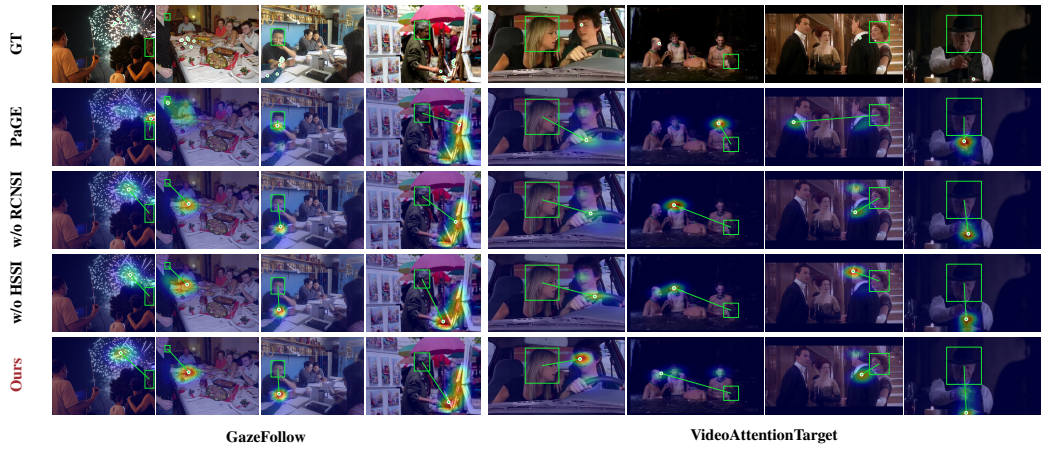

Figure 4: Qualitative component analysis on GazeFollow and VideoAttentionTarget. Rows show GT, PaGE, w/o RCNSI, w/o HSSI, and Ours. Each column uses the same image and queried person. Green boxes mark the queried head; green points show annotations in the GT row and predicted peaks in the other rows. Lines connect head centers to predicted peaks. The four predicted heatmaps in each column share a color scale.

Table 2: Component ablation. Base uses the partially shared backbone and three interaction layers with both modules disabled. HSSI-only and RCNSI-only each enable the named module on Base; Ours enables both. All configurations were trained independently. Parameters are in millions. Bold red and underline mark the best and second-best values; ties share a rank.

		GazeFollow			VideoAttentionTarget		
Model	Params ↓	AUC ↑	Avg L_2 ↓	Min L_2 ↓	AUC ↑	L_2 ↓	AP ↑
PaGE (Ye et al., 2026)	187.306	<u>0.9626</u>	0.0887	0.0338	0.9595	0.0807	0.9378
Base	153.094	0.9622	0.0890	0.0344	0.9597	0.0827	**0.9439**
+ HSSI	153.300	<u>0.9626</u>	<u>0.0856</u>	<u>0.0322</u>	<u>0.9610</u>	0.0814	0.9290
+ RCNSI	153.295	<u>0.9626</u>	0.0872	0.0329	0.9608	<u>0.0793</u>	0.9375
Ours (both)	153.501	**0.9630**	**0.0851**	**0.0317**	**0.9621**	**0.0785**	<u>0.9389</u>

aggregation through reliability-weighted routing, while the spatial gate controls where this information corrects the heatmap logits. The component ablations and within-model comparison show that pre-decoding feature modulation and post-decoding heatmap refinement jointly improve gaze target localization.

Limitations. The current model uses single frames and depends on head-box quality. Candidate-based heatmap refinement also depends on the locations selected from the coarse heatmap. If the candidate set does not cover the target region, scene anchors still provide broader scene context, but no candidate feature is sampled directly from the target. Temporal context and more flexible candidate selection are useful directions for future work.

6 Conclusion

To address gaze ambiguity arising from weak target cues before decoding and competing heatmap peaks afterward, we proposed SPECTRA-Gaze. HSSI modulates scene features using head cues, while RCNSI uses candidate reliability to refine the coarse heatmap. Experiments on GazeFollow and VideoAttentionTarget showed improved localization over strong published baselines. Under the same training and evaluation protocol, SPECTRA-Gaze outperformed PaGE ViT-B without distillation across all six metrics, with 18.05% fewer parameters.

AI USE STATEMENT

We used generative AI tools for English-language editing, LaTeX formatting, and assistance with literature searches. The authors are responsible for the scientific claims, experimental results, references, and final manuscript.

ETHICS STATEMENT

This study used public gaze benchmarks and collected no new human-subject data. Gaze target estimation processes head and scene cues associated with behavioral attention and raises privacy concerns in surveillance and monitoring applications. Its outputs estimate gaze locations rather than a person's intent or mental state. Deployment requires respect for dataset licenses, privacy, and consent, together with assessment of demographic differences and head-detection failures.

REPRODUCIBILITY STATEMENT

Section 3 specifies the model and losses, and Appendix A documents data processing, optimization, evaluation coordinates, and checkpoint selection. Appendices B and C detail the two modules. We fixed the evaluation checkpoint before training and trained each ablation configuration independently. We will release the code, configurations, and model weights upon acceptance.

REFERENCES

Jun Bao, Buyu Liu, and Jun Yu. ESCNet: Gaze target detection with the understanding of 3D scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14126–14135, 2022.

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 (CVPR)*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Cao_Gaze_Target_Estimation_Anywhere_with_Concepts_CVPR_2026_paper.pdf.

Lu Chi, Borui Jiang, and Yadong Mu. Fast Fourier convolution. In *Advances in Neural Information Processing Systems*, volume 33, pp. 4479–4488, 2020.

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 *Computer Vision – ECCV 2018*, volume 11209 of *Lecture Notes in Computer Science*, pp. 397–412, 2018. doi: 10.1007/978-3-030-01228-1_24. URL https://doi.org/10.1007/978-3-030-01228-1_24.

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 (CVPR)*, pp. 5395–5405, 2020.

Yi Fang, Jiapeng Tang, Wang Shen, Wei Shen, Xiao Gu, Li Song, and Guangtao Zhai. Dual attention guided gaze target detection in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11390–11399, 2021.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Proceedings of the First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=tEYskw1VY2.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=uYLFoz1vlAC.

John Guibas, Morteza Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, and Bryan Catanzaro. Efficient token mixing for transformers via adaptive Fourier neural operators. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=EXHG-A3jlM.

Anshul Gupta, Samy Tafasca, and Jean-Marc Odobez. A modular multimodal architecture for gaze target prediction: Application to privacy-sensitive settings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 5041–5050, 2022.

Anshul Gupta, Samy Tafasca, Naravich Chutisilp, and Jean-Marc Odobez. A unified model for gaze following and social gaze prediction. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–9, 2024a. doi: 10.1109/FG59268.2024.10581955.

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*, volume 37, pp. 15646–15673, 2024b.

Yuqi Hou, Zhongqun Zhang, Nora Horanyi, Jaewon Moon, Yihua Cheng, and Hyung Jin Chang. Multi-modal gaze following in conversational scenarios. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1186–1195, 2024.

Tianlei Jin, Qizhi Yu, Shiqiang Zhu, Zheyuan Lin, Jie Ren, Yuanhai Zhou, and Wei Song. Depth-aware gaze-following via auxiliary networks for robotics. *Engineering Applications of Artificial Intelligence*, 113:104924, 2022. doi: 10.1016/j.engappai.2022.104924.

Yanzhao Li, Jin Li, Xiaona Zhang, and Hong Liang. Uncertainty-aware general gaze following via circular direction distribution learning and probabilistic gaze geometry modeling. *Journal of Eye Movement Research*, 19(4):88, 2026. doi: 10.3390/jemr19040088.

Dongze Lian, Zehao Yu, and Shenghua Gao. Believe it or not, we know what you are looking at! In *Computer Vision – ACCV 2018*, volume 11363 of *Lecture Notes in Computer Science*, pp. 35–50. Springer International Publishing, 2019. doi: 10.1007/978-3-030-20893-6_3. URL https://arxiv.org/abs/1907.02364.

Feiyang Liu, Dan Guo, Jingyuan Xu, Zihao He, Shengeng Tang, Kun Li, and Meng Wang. Towards pixel-level prediction for gaze following: Benchmark and approach. *arXiv preprint arXiv:2412.00309*, 2024a. doi: 10.48550/arXiv.2412.00309.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. VMamba: Visual state space model. In *Advances in Neural Information Processing Systems*, volume 37, pp. 103031–103063, 2024b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Athul M. Mathew, Haithem Hermassi, Thariq Khalid, Arshad Ali Khan, and Riad Souissi. GazeVLM: A vision-language model for multi-task gaze understanding. *arXiv preprint arXiv:2511.06348*, 2025. doi: 10.48550/arXiv.2511.06348. URL https://arxiv.org/abs/2511.06348.

Jiajie Mi, Xinyu Liu, Mengke Song, and Chenglizhao Chen. Multi-scale object-aware gaze estimation via geometric reasoning. *arXiv preprint arXiv:2606.29334*, 2026. doi: 10.48550/arXiv.2606.29334. URL https://arxiv.org/abs/2606.29334.

Qiaomu Miao, Minh Hoai, and Dimitris Samaras. Patch-level gaze distribution prediction for gaze following. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 880–889, 2023.

Qiaomu Miao, Alexandros Graikos, Jingwei Zhang, Sounak Mondal, Minh Hoai, and Dimitris Samaras. Diffusion-refined VQA annotations for semi-supervised gaze following. In *Computer Vision – ECCV 2024*, pp. 439–457. Springer, 2024. doi: 10.1007/978-3-031-72933-1_25.

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 (ICCV)*, pp. 5371–5381, 2025.

Álvaro Nieva-Suárez, Marta Marrón-Romera, Cristina Losada-Gutiérrez, and Irene Guardiola-Luna. Towards fusing gaze estimation and object prediction: What are you looking at? *Engineering Applications of Artificial Intelligence*, 157:111113, 2025. doi: 10.1016/j.engappai.2025.111113.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 3942–3951, 2018. doi: 10.1609/aaai.v32i1.11671.

Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. FcaNet: Frequency channel attention networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 783–792, 2021.

Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. In *Advances in Neural Information Processing Systems*, volume 34, pp. 980–993, 2021.

Adrià Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? In *Advances in Neural Information Processing Systems*, volume 28, pp. 199–207. Curran Associates, Inc., 2015. URL https://papers.neurips.cc/paper_files/paper/2015/hash/ec8956637a99787bd197eacd77acce5e-Abstract.html.

Adrià Recasens, Carl Vondrick, Aditya Khosla, and Antonio Torralba. Following gaze in video. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1444–1452, 2017. doi: 10.1109/ICCV.2017.160.

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

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 Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1586–1594. International Joint Conferences on Artificial Intelligence Organization, 2026. doi: 10.24963/ijcai.2026/177.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. arXiv:2508.10104, 2025. URL https://arxiv.org/abs/2508.10104.

Yuehao Song, Xinggang Wang, Jingfeng Yao, Wenyu Liu, Jinglin Zhang, and Xiangmin Xu. ViTGaze: Gaze following with interaction features in vision transformers. *Visual Intelligence*, 2:31, 2024. doi: 10.1007/s44267-024-00064-9.

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 (ICCV)*, pp. 20935–20946, 2023.

Samy Tafasca, Anshul Gupta, Victor Bros, and Jean-Marc Odobez. Toward semantic gaze target detection. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. doi: 10.52202/079017-3858.

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 (CVPR)*, pp. 2008–2017, 2024b.

Yuki Tatsunami and Masato Taki. FFT-based dynamic token mixer for vision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15328–15336, 2024. doi: 10.1609/aaai.v38i14.29457.

Francesco Tonini, Cigdem Beyan, and Elisa Ricci. Multimodal across domains gaze target detection. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pp. 420–431. Association for Computing Machinery, 2022. doi: 10.1145/3536221.3556624.

Francesco Tonini, Nicola Dall'Asen, Cigdem Beyan, and Elisa Ricci. Object-aware gaze target detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21860–21869, 2023.

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7794–7803, 2018.

Yaokun Yang and Feng Lu. GazeLLM: A plug-and-play zero-shot LLM reasoning framework for boosting gaze target detection. *Visual Intelligence*, 3:26, 2025. doi: 10.1007/s44267-025-00101-1.

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:2607.04860, 2026. URL https://arxiv.org/abs/2607.04860.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision Mamba: Efficient visual representation learning with bidirectional state space model. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62429–62442, 2024. URL https://proceedings.mlr.press/v235/zhu24f.html.

SUPPLEMENTARY MATERIAL

This supplement specifies the training protocol, tensor dimensions, and implementation details omitted from the main text. Notation follows Section 3; supplementary equations use the prefix S. The implementation described here corresponds to the scalar-gated full model used in the reported experiments.

A TRAINING AND IMPLEMENTATION DETAILS

A.1 DATA PROCESSING AND OPTIMIZATION

We jointly trained a single model for each configuration on the combined GazeFollow and VideoAttentionTarget (VAT) training sets. Samples from both datasets were shuffled together in each training epoch. The same fixed checkpoint was evaluated separately on the two test sets; no dataset-specific fine-tuning was applied. Training used neither ChildPlay data nor unlabeled task-distillation data. GazeFollow used every training record, whereas VAT used every third training frame. Random cropping, horizontal flipping, and head-box jitter were each applied with probability 0.5. Scene images and head crops were resized to 512×512 and 256×256, respectively, using the backbone transforms. Training heatmaps had resolution 64×64, with a Gaussian target of standard deviation three cells and unit maximum.

Table 3 summarizes the two-stage schedule. Stage 2 loaded the final Stage 1 checkpoint, then fine-tuned the backbone together with the task-specific modules. Optimization used non-fused AdamW (Loshchilov & Hutter, 2019), cosine learning-rate decay, bfloat16 automatic mixed precision, and gradient clipping at norm one. The learning-rate floor was 10^{-7} in both stages. DINOv3 feature dropout preceded projection, with rates 0.1 and 0.5 in Stages 1 and 2, respectively. Interaction blocks used stochastic depth with rate 0.1.

Hardware and software. Each training and evaluation run used one NVIDIA GeForce RTX 4090 GPU. The evaluation environment used PyTorch 2.12.1 with the CUDA 13.0 build and bfloat16 automatic mixed precision. Training used batch sizes of 32 and 8 in Stages 1 and 2, respectively, without gradient accumulation for the reported ViT-B configurations. Benchmark evaluation used batch size 16.

Table 3: Two-stage optimization settings. Batch size is the number of training samples per optimizer update for these model sizes.

Setting	Stage 1	Stage 2
Trainable components	Task-specific modules	Backbone and task-specific modules
Epochs	15	5
Batch size	32	8
Base learning rate	10^{-3}	10^{-5}
Weight-decay setting	5×10^{-2}	10^{-2}
Warm-up iterations	100	500
Warm-up initial rate	10^{-4}	10^{-7}
DINOv3 feature dropout	0.1	0.5

Training of the single-module ablations. We constructed HSSI-only and RCNSI-only configurations by enabling the respective module on the compact base network while keeping the other module disabled. Each configuration was trained and evaluated independently. These comparisons share data processing and the two-stage training schedule. Removing RCNSI also removes its auxiliary reliability loss, so the comparison includes both changes.

A.2 Supervision and coordinate conventions

Heatmap binary cross-entropy was evaluated only on in-frame samples, while in/out binary cross-entropy used all samples. Both losses operated directly on logits with mean reduction. For candidate k, let c_k be its endpoint-inclusive 32×32 grid coordinate and y the target-heatmap argmax coordinate divided by 63. The operator sg stops gradients, $\|\cdot\|_2$ is the Euclidean norm, and $\sigma_r = 0.08$ sets the Gaussian width. The soft reliability target a_k^{tar} is

$$a_k^{\mathrm{tar}} = \exp\left\{-\frac{\|\mathrm{sg}(c_k) - y\|_2^2}{2\sigma_r^2}\right\}, \qquad \sigma_r = 0.08. \tag{S1}$$

Binary cross-entropy between ρ_k and a_k^{tar}, restricted to in-frame samples, defines $\mathcal{L}_{\mathrm{rel}}$. With RCNSI enabled, the total loss follows Equation equation 7. HSSI requires neither frequency-band annotations nor an auxiliary loss.

Training reliability targets and evaluation coordinates use different conventions in the implementation. Candidate coordinates divide grid indices by 31, and reliability targets divide heatmap indices by 63. Evaluation instead divides the predicted heatmap argmax indices by 64, without a half-cell offset. The PaGE reproduction, Base, and all module configurations used this same evaluation coordinate convention.

A.3 Backbone sharing and tensor dimensions

The compact encoder used DINOv3 ViT-B/16 pretrained on LVD-1689M (Siméoni et al., 2025). The scene and head branches shared the embedding module, rotary-position module, and first four transformer blocks. Their upper eight blocks and final normalization remained independent. The shared rotary-position module constructs position coordinates dynamically for each input resolution. Both streams use the shared parameters in separate forward passes.

The projected scene and head features are S and Q. HSSI updates S before the head-position prompt and cross-stream interaction stack. Each stream contains four register tokens and one in/out token in addition to the spatial tokens. The two in/out tokens jointly predict in-frame probability, whereas RCNSI modifies only the heatmap logits after decoding. Table 4 fixes the notation used throughout this supplement.

Table 4: Tensor specifications, omitting the batch dimension. Spatial tokens exclude the four register tokens and one in/out token in each stream.

Symbol	Meaning	Shape
S, Q	Projected scene and head features	$256 \times 32 \times 32$; $256 \times 16 \times 16$
$\widehat{F}_s, \widehat{F}_h$	HSSI scene and head spectra	$96 \times 32 \times 32$; $96 \times 16 \times 16$
M_k	One HSSI frequency-band mask	32×32
B	HSSI band descriptors	8×96
$\{x_i\}_{i=1}^{1024}, h$	RCNSI receivers on the 32×32 grid and pooled head feature	1024×96; 96
$\{v_j\}_{j=1}^{24}$	Eight candidate and 16 anchor features	24×96
$L, L^\star, H$	Coarse logits, refined logits, heatmap	64×64 each

Parameter accounting. The compact base contains 153,093,889 unique parameters, compared with 187.306M in the PaGE reproduction. HSSI adds 205,769 parameters, and RCNSI adds 201,350, giving 153,501,008 for the full model. The total reduction relative to our PaGE ViT-B reproduction is approximately 18.05%.

A.4 Benchmark metrics and checkpoint loading

GazeFollow training uses one gaze point per record, while testing retains multiple annotations (Recasens et al., 2015). Its AUC compares the heatmap, bilinearly resized to the original image dimensions, with a binary map of annotated gaze pixels. Average L_2 measures distance to the mean annotation; minimum L_2 measures distance to the closest annotation. VAT AUC and L_2 use in-frame samples, while in/out AP uses all retained samples (Chong et al., 2020). The VAT AUC implementation constructs a binary rectangular tolerance region on the 64×64 grid, retaining the PaGE boundary conventions. Final VAT evaluation uses every test frame; sparse evaluation during training is a separate monitoring procedure. PaGE, HSSI-only, RCNSI-only, and the full model were evaluated with batch size 16 and bfloat16 automatic mixed precision. RCNSI used a 32×32 grid and eight candidates.

We fixed Stage 2 epoch 4 as the evaluation checkpoint before training for every local configuration. The reported full model used scalar residual gates. Loading reconciles the two supported DINOv3 block-key layouts. Missing keys, unexpected keys, and tensor shapes must be checked, including the backbone, to ensure that all weights are restored. Each run was evaluated using a single checkpoint, without checkpoint ensembling.

A.5 Qualitative predictions and internal states

Figure 4 uses predictions from independently trained PaGE, HSSI-only, RCNSI-only, and full-model checkpoints. We export the heatmaps with evaluation batch size 16 and the 32×32 internal grid. Within each example, all four heatmaps share the same maximum for color normalization and are overlaid on the original image with opacity 0.4. The predicted point is the heatmap maximum, and the original image aspect ratio and full field of view are retained.

We exported Figure 5 from the fixed full-model checkpoint in evaluation mode, using the first three queries in the unshuffled GazeFollow test loader. The RCNSI instance was explicitly evaluated on a 32×32 grid. Scene activation is the channel-wise mean of the absolute post-interaction scene features passed to RCNSI, giving a 32×32 map. HSSI band dominance is $d(u) = \max_{1 \le k \le 8} M_k(u)$, computed directly from the eight soft band masks before any spatial reconstruction. Since $\sum_k M_k(u) = 1$, dominance lies in $[1/8, 1]$ and measures the concentration of frequency-band assignment. The horizontal and vertical DCT-bin indices are divided by 31; these axes describe frequency rather than image position. The remaining maps use image coordinates, increasing rightward and downward. The displayed spatial gate is $\mathcal{U}(T)$, where $T_i = \tau_i e_i \sqrt{u_i}$, and the correction is the applied residual in Equation (6); both maps have resolution 64×64. We used shared color limits across the three rows: $[0, 4.14005]$ for activation, $[0.125, 1]$ for band dominance, $[0, 1]$ for the gate, and $[-0.24542, 0.24542]$ for the logit correction. The correction scale is symmetric around zero, with red indicating increased logits and blue indicating decreased logits.

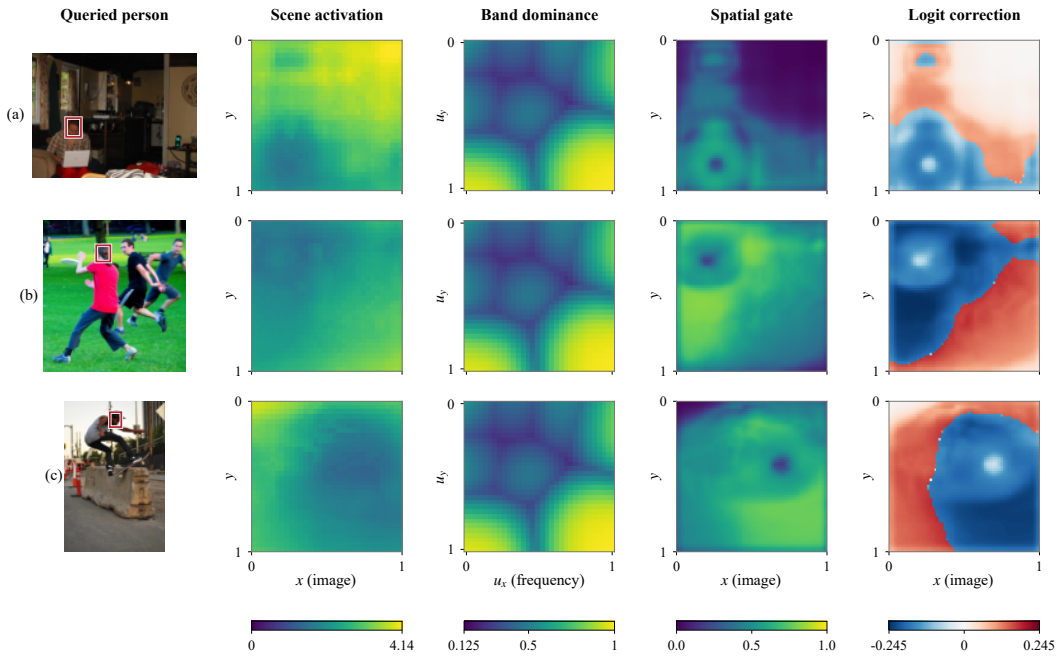

Figure 5: Internal states for the first three GazeFollow test queries. Red boxes mark the queried heads. Columns show channel-mean absolute scene activation, HSSI band dominance $\max_k M_k(u)$, the upsampled spatial gate $\mathcal{U}(T)$, and signed logit corrections $L^\star - L$. Each column shares one color scale across rows. Frequency coordinates (u_x, u_y) are normalized DCT-bin indices, with DC at the upper left; (x, y) are normalized image coordinates. Appendix A.5 gives the extraction details.

B HSSI: Detailed Formulation

HSSI constructs frequency bands from joint scene and head descriptors, then uses interactions among band descriptors to modulate scene processing. This section details the mask construction and implementation constants used in Section 3.2.

B.1 Spectral projection and band geometry

Separate 1×1 convolutions project S and Q to 96 channels. The implementation casts their outputs to float32 before passing them to the separable orthonormal DCT routine. For each spatial size n, the routine constructs the cosine basis with the usual $\sqrt{2/n}$ factor and rescales its first row by $1/\sqrt{2}$. The inverse uses the transposed orthonormal bases in reversed order.

Global averages of $\log(1 + |\widehat{F}_s|)$ and $\log(1 + |\widehat{F}_h|)$ form a 192-dimensional descriptor. An MLP with widths $192 \to 96 \to 40$ and a GELU activation predicts eight sets of five band parameters. Equation (1) converts these outputs into centers μ_k, scales d_k, and shear coefficients o_k. Its constants are $\alpha_c = 0.5$, $d_{\min} = 0.06$, $d_{\max} = 0.40$, and $o_{\max} = 0.15$. Reference centers are clamped to $[0.01, 0.99]$ before the logit transform. Their fixed values are $(0.08, 0.08)$, $(0.30, 0.08)$, $(0.60, 0.08)$, $(0.90, 0.15)$, $(0.08, 0.35)$, $(0.35, 0.40)$, $(0.15, 0.85)$, and $(0.75, 0.75)$. Thus, adaptive centers retain a bounded dependence on these reference locations. At normalized frequency coordinate $u \in [0, 1]^2$, let A_k be lower triangular. Its diagonal contains the two components $(d_{k,x}, d_{k,y})$ of d_k, and its lower-left entry is o_k. In the normalization below, ℓ indexes the eight bands and $\|\cdot\|_2$ is the Euclidean norm. The normalized anisotropic Gaussian masks are

$$M_k(u) = \frac{\exp\left\{-\frac{1}{2}\left\|A_k^{-1}(u - \mu_k)\right\|_2^2\right\}}{\sum_{\ell=1}^{8} \exp\left\{-\frac{1}{2}\left\|A_\ell^{-1}(u - \mu_\ell)\right\|_2^2\right\}}. \tag{S2}$$

Both coordinate axes include their endpoints, and the masks sum to one across bands at every scene-frequency location. The head spectrum conditions band geometry, while the spectral scan operates on the scene spectrum.

B.2 Band descriptors and modulation order

Let $E_s(u) = \log(1 + |\widehat{F}_s(u)|)$ be the scene log-magnitude at frequency u. Let f_{geo} be a learned affine map from five geometry parameters to 96 channels. The initial band descriptor z_k is

$$z_k = \frac{\sum_u M_k(u) E_s(u)}{\max\left\{\sum_u M_k(u), 10^{-6}\right\}} + f_{\mathrm{geo}}([\mu_k; d_k; o_k]) \,. \tag{S3}$$

Here, $\sum_u$ runs over frequency locations, semicolons denote concatenation, and 10^{-6} prevents a zero denominator. Stack the eight descriptors z_k as rows of B. Let LN denote layer normalization and $\mathcal{A}$ four-head self-attention. The updated matrix is $\widetilde{B} = B + \mathcal{A}(\mathrm{LN}(B))$, whose row k is $\widetilde{z}_k$. The residual therefore bypasses layer normalization.

A $96 \to 192$ affine map produces the pre- and post-scan states for each band. The implementation applies $\tanh$ to each band state before combining states with the masks, as in Equation (2). The two resulting modulation maps have 96 channels on the 32×32 scene-frequency grid.

B.3 Selective scanning and residual writeback

Equations (3) and (4) specify the spectral gate and scene-feature update. The scan input uses signed logarithmic compression and pre-modulation with strength 0.25. The scan output is bounded by $0.5 \tanh(\cdot)$, multiplied by the original scene coefficients, and post-modulated with strength 0.25. An inverse DCT and a bias-free 1×1 convolution restore the 256-channel scene representation. The output residual is cast to the scene-feature dtype before addition.

Each SS2D adapter applies one-group normalization before the project-provided selective-scan implementation (Liu et al., 2024b). Its settings are channel width 96, state size 16, expansion ratio one, rank ratio one, convolution width three, and forward mode `v2`. All reported experiments use the `mamba_yolo` backend for selective scanning.

Bounds and computational scope. The scalar gate $g_s = \tanh(\gamma)$ starts at zero for a newly constructed module. This initializes HSSI as an identity on its projected scene input. The spectral multipliers are bounded, while the learned write projection is unconstrained and determines the scale of the spatial feature update. Self-attention operates on eight band descriptors, and selective scanning operates on the scene-frequency grid. The DCT routine uses dense separable matrix products.

B.4 Fixed-weight HSSI interventions on GazeFollow

We evaluated three interventions in the trained Full model on all 4,782 GazeFollow test queries, using batch size 16, bfloat16 inference, and a 32×32 RCNSI grid with eight candidates. All model weights and the remaining computation were held fixed. Head-descriptor mismatch replaces the head spectral-energy descriptor used for band prediction with that of another query through a fixed permutation without self-matches; the downstream head stream is retained. Spatial-path replacement substitutes projected spatial features for the scene DCT coefficients, matches their per-channel root-mean-square magnitude, and omits the inverse DCT on writeback. Uniform band masks assign each frequency equally to the eight bands, retaining the remaining band-processing operations.

Table 5 shows the largest increase in average localization error for head-descriptor mismatch, followed by spatial-path replacement and uniform band masks.

Table 5: Fixed-weight HSSI interventions on GazeFollow. Δ is the change in average L_2 relative to Full under the same evaluation settings.

Configuration	Avg $L_2 \downarrow$	Δ Avg L_2
Full	0.08507596	—
Head-descriptor mismatch	0.08538927	+0.00031331
Spatial-path replacement	0.08533539	+0.00025943
Uniform band masks	0.08519845	+0.00012248

C RCNSI: Detailed Formulation

RCNSI aggregates candidate features and scene anchors, then uses the aggregated context to refine the coarse logits. This section specifies discrete selection, reliability inputs, routing, and the coordinate convention used by the evaluated implementation.

C.1 Candidate selection and local statistics

A 1×1 projection produces the $96 \times 32 \times 32$ scene grid with receiver features x_i. The coarse logits L are independently average-pooled to 32×32 and normalized by a spatial softmax to obtain p. This auxiliary distribution is distinct from the final sigmoid heatmap. Let $\mathcal{P}$ contain positions equal to their 3×3 max-pooled value. Let $\widetilde{p}_i$ be the adjusted selection score. The operator $\mathrm{TopK}(\cdot, 8)$ returns the eight highest-scoring indices, collected in $\mathcal{I}$:

$$\mathcal{I} = \mathrm{TopK}(\widetilde{p}, 8), \qquad \widetilde{p}_i = \begin{cases} p_i, & i \in \mathcal{P}, \\ p_i - 1, & i \notin \mathcal{P}. \end{cases} \tag{S4}$$

Candidate features are indexed directly from the scene grid at $\mathcal{I}$. The implementation has no separate validity mask: non-maximal positions can fill unused slots when fewer than eight maxima exist. Equal-valued plateau maxima are resolved by the underlying top-k ordering.

Each candidate uses a zero-padded 3×3 probability neighborhood. Local mass is the sum of its nine values; normalization divides these values by the mass clamped below at 10^{-8}. Entropy uses the normalized probabilities, clamps the logarithm argument below at 10^{-8}, and divides by $\log 9$. Peak margin is the difference between the two largest neighborhood values divided by the largest value clamped below at 10^{-8}. A pooled head feature is projected from 256 to 96 dimensions. The reliability MLP concatenates candidate appearance, head appearance, mass, entropy, peak margin, and two displacement coordinates, giving 197 inputs. Its widths are $197 \to 64 \to 1$, with GELU between the affine layers and sigmoid applied to the output logit ρ_k. All inputs to this MLP are detached; its parameters still receive gradients through the task and auxiliary reliability losses.

C.2 Head-box displacement

RCNSI uses the same 32×32 spatial grid as the projected scene features and head-box coordinates. Candidate coordinates c_k lie on an endpoint-inclusive grid with spacing $1/31$. The head-box routine rounds normalized box corners onto the scene grid. The tuple $(r_y^-, r_x^-, r_y^+, r_x^+)$ gives the top, left, bottom, and right grid indices. Let $\widetilde{c}_h$ be the normalized head-box center after coordinate-wise clipping to $[0, 1]$. The displacement from this center to candidate k is δ_k:

$$\widetilde{c}_h = \mathrm{clip}_{[0,1]}\left(\begin{bmatrix} (r_x^- + r_x^+)/(2 \cdot 32) \\ (r_y^- + r_y^+)/(2 \cdot 32) \end{bmatrix} \right), \qquad \delta_k = c_k - \widetilde{c}_h. \tag{S5}$$

The displacement enters the candidate-reliability MLP, while receiver–memory coordinate differences enter the routing geometry term.

C.3 Memory construction and rejectable routing

Memory contains eight candidate features and 16 anchors from 4×4 adaptive average pooling of the scene grid. Anchor coordinates are assigned on an endpoint-inclusive 4×4 grid. Candidate prior

weights equal their predicted reliabilities, while anchor prior weights are fixed at 0.25. Queries project $x_i + h$, keys project memory features v_j, and both projections are ℓ_2-normalized. Values use a separate bias-free linear projection.

Equation (5) combines feature similarity, relative geometry, and source reliability. The similarity multiplier is four, and the geometry MLP has widths $2 \to 32 \to 1$ with GELU. Source reliability enters as $\log(\max\{s_j, 10^{-8}\})$, retaining the implementation's lower clamp. A learnable null logit b_{null}, initialized to two, enters the normalization. Let m index the 24 real sources. The routing weight r_{ij} for receiver i and source j is

$$r_{ij} = \frac{\exp(\ell_{ij})}{\exp(b_{\text{null}}) + \sum_{m=1}^{24} \exp(\ell_{im})}, \qquad j = 1, \ldots, 24. \tag{S6}$$

The remaining probability is assigned to a zero-valued null source. The message $m_i = \sum_{j=1}^{24} r_{ij} W_v v_j$ is added to x_i, then reshaped to the spatial grid. A residual SS2D adapter updates this aggregate, followed by a 3×3 convolution from 96 to 64 channels, GELU, and a 1×1 scalar-output convolution. Bilinear upsampling with `align_corners=False` produces the 64×64 correction map R. Routing uses 1,024 receivers on the 32×32 grid and 24 real sources, plus the null source.

C.4 Spatial gate, gradients, and correction bounds

Receiver entropy e_i uses the same local normalization as candidate entropy. Reliable support is $u_i = \sum_{k=1}^{8} r_{ik} a_k$, using only candidate sources rather than the anchors. The target-gate MLP receives x_i, e_i, u_i, the null probability, and p_i, totaling 100 inputs. It has widths $100 \to 32 \to 1$, with GELU and an output sigmoid. Its final layer starts with zero weights and bias -1.5, giving an initial learned factor $\tau_i = \sigma(-1.5)$. The factor is learned during training, with no monotonicity constraint.

Inputs to this MLP are detached, but entropy and reliable support also multiply its output through $T_i = \tau_i e_i \sqrt{\max(u_i, 0)}$. Those external factors retain their gradient paths. The implementation bilinearly upsamples T and subtracts the spatial mean from R before applying Equation (6). The residual scale is $\beta = 0.5$, and the scalar gate is $g_r = \tanh(\eta)$, with zero initialization for a new module. The per-pixel correction magnitude is bounded by 0.5 under the stated gate ranges. Mean-centering occurs before the nonlinearities and spatial weighting; the final applied correction need not have zero spatial mean.

Entropy quantifies local heatmap dispersion, and candidate reliability is trained against a target-proximity score. The null route regulates the contribution of aggregated context, while the spatial gate controls its effect on the logit update. RCNSI operates on the heatmap branch; the in/out head uses the in/out tokens from the scene and head streams.

D Architectural Positioning and Comparison Scope

Gaze-LLE combines pretrained scene features with head-position prompting (Ryan et al., 2025). PaGE models scene–head interactions with a dual-stream architecture (Ye et al., 2026). Our compact core shares part of the encoder and uses three bidirectional interaction layers, while retaining separate scene and head computations. These changes account for the parameter reduction; HSSI and RCNSI add parameters to that compact core.

HSSI modulates the scene representation before interaction, whereas RCNSI refines logits after coarse prediction. The single-module configurations enable each module separately; RCNSI configurations also use the auxiliary reliability loss described in Appendix A.

Tables 1 and 2 report benchmark performance and parameter counts. The local PaGE comparison uses the shared training and evaluation protocol, while published results retain their original protocols.

D.1 WITHIN-MODEL HEATMAP REFINEMENT

We evaluated the same fixed SPECTRA-Gaze checkpoint on all 4,782 GazeFollow test queries and 20,655 in-frame VideoAttentionTarget test queries. Coarse and refined heatmaps use the same localization decoder.

Table 6: Localization before and after RCNSI refinement. Δ is refined minus coarse; negative values indicate lower error.

Metric	Coarse	Refined	Δ (95% interval)
GF Avg $L_2 \downarrow$	0.085429	**0.085076**	-0.000353 $[-0.000557, -0.000202]$
GF Min $L_2 \downarrow$	0.031937	**0.031742**	-0.000196 $[-0.000293, -0.000102]$
VAT $L_2 \downarrow$	0.078735	**0.078543**	-0.000192 $[-0.000313, -0.000077]$

Intervals use 2,000 paired bootstrap resamples of GazeFollow images and VideoAttentionTarget videos. GazeFollow AUC changed from 0.963002 to 0.962975, while VideoAttentionTarget AUC rose from 0.961891 to 0.962068. The in/out prediction is shared by both outputs, so VideoAttentionTarget AP is unchanged.

D.2 PAIRED LOCALIZATION COMPARISON ON GAZEFOLLOW

We compared SPECTRA-Gaze with our PaGE reproduction on all 4,782 GazeFollow test queries using their fixed checkpoints. Both models were evaluated with identical preprocessing, coordinate decoding, batch size 16, and bfloat16 inference. We paired predictions by test-image and query identity and drew 2,000 bootstrap resamples of test images with replacement, retaining the paired predictions within each resample. The 2.5th and 97.5th percentiles of the mean error difference give the 95% confidence interval for each metric, quantifying test-sample uncertainty for these checkpoints.

Table 7: Paired localization results on GazeFollow. Δ is SPECTRA-Gaze minus PaGE; negative values indicate lower error. Intervals are computed separately for each metric.

Metric	PaGE	SPECTRA-Gaze	Δ	95% interval
Avg $L_2 \downarrow$	0.088725	0.085076	-0.003649	$[-0.005561, -0.001728]$
Min $L_2 \downarrow$	0.033849	0.031742	-0.002108	$[-0.003565, -0.000656]$

SPECTRA-Gaze reduces both average and minimum localization error in this paired comparison (Table 7). Both confidence intervals lie below zero, supporting the GazeFollow localization improvement over the PaGE reproduction.

D.3 PERFORMANCE ACROSS TRAINING SEEDS

Table 8 summarizes three independently trained runs of PaGE and SPECTRA-Gaze. We report the mean and sample standard deviation across runs ($n = 3$). The single-run results in Table 1 are included in this summary.

Table 8: Performance across three training seeds (mean $\pm$ sample standard deviation).

	GazeFollow			VideoAttentionTarget		
Model	AUC $\uparrow$	Avg $L_2 \downarrow$	Min $L_2 \downarrow$	AUC $\uparrow$	$L_2 \downarrow$	AP $\uparrow$
PaGE	0.9626 ± 0.0001	0.0882 ± 0.0006	0.0338 ± 0.0001	0.9611 ± 0.0014	0.0801 ± 0.0010	0.9293 ± 0.0095
SPECTRA-Gaze	0.9630 ± 0.0001	0.0853 ± 0.0002	0.0318 ± 0.0001	0.9620 ± 0.0002	0.0788 ± 0.0005	0.9383 ± 0.0010

E CANDIDATE RELIABILITY ANALYSIS

We examined whether RCNSI reliability scores distinguish target-relevant candidates using the fixed full-model checkpoint with batch size 16 and a 32×32 RCNSI grid. The analysis included all eight candidates from each of 4,782 GazeFollow and 20,655 VAT in-frame test queries, yielding 38,256 and 165,240 candidate-level observations, respectively. We grouped predicted reliability into ten fixed intervals of width 0.1.

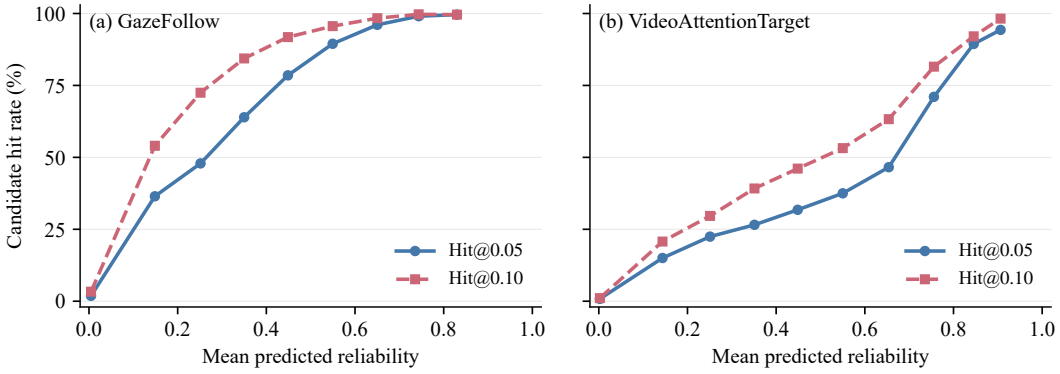

Figure 6: Candidate quality across reliability intervals. Each point uses the mean predicted reliability of one nonempty interval. Hit@δ is the percentage of candidates within normalized Euclidean distance δ of any GazeFollow annotation or the single VAT target. Solid and dashed curves use $\delta = 0.05$ and 0.10, respectively. Table 9 reports all interval counts.

Table 9 additionally reports mean predicted reliability and the mean soft proximity target from Equation (S1). For this diagnostic, the target coordinate is the peak of the Gaussian target map on VAT and of the average Gaussian annotation map on GazeFollow, with peak indices divided by 63. The Gaussian scale is $\sigma_r = 0.08$. This soft target measures graded proximity, while Hit@δ measures localization within a fixed tolerance.

Table 9: Candidate reliability by score interval. N is the number of candidates, $\bar{a}$ their mean predicted reliability, and $\bar{a}^{\mathrm{tar}}$ their mean Gaussian proximity target. Hit@0.05 is a percentage. All ten intervals are shown; an empty interval has no estimated means.

	GazeFollow				VideoAttentionTarget			
Score interval	N	$\bar{a}$	$\bar{a}^{\mathrm{tar}}$	Hit@0.05	N	$\bar{a}$	$\bar{a}^{\mathrm{tar}}$	Hit@0.05
$[0.0, 0.1)$	29,856	0.005	0.006	1.87	138,984	0.002	0.011	0.83
$[0.1, 0.2)$	1,042	0.149	0.165	36.47	2,131	0.144	0.208	15.02
$[0.2, 0.3)$	977	0.252	0.313	47.90	1,501	0.251	0.296	22.45
$[0.3, 0.4)$	1,159	0.350	0.457	63.93	1,725	0.351	0.366	26.55
$[0.4, 0.5)$	1,219	0.449	0.596	78.51	1,879	0.449	0.433	31.77
$[0.5, 0.6)$	1,207	0.550	0.719	89.48	1,917	0.550	0.503	37.51
$[0.6, 0.7)$	1,476	0.650	0.852	96.07	2,718	0.654	0.583	46.58
$[0.7, 0.8)$	1,068	0.744	0.925	99.06	4,796	0.756	0.760	71.04
$[0.8, 0.9)$	252	0.830	0.955	99.60	9,308	0.846	0.883	89.42
$[0.9, 1.0]$	0	—	—	—	281	0.906	0.951	94.31

Hit@0.05 increases across the occupied reliability intervals on both datasets. Between the $[0.0, 0.1)$ and $[0.7, 0.8)$ intervals, it rises from 1.87% to 99.06% on GazeFollow and from 0.83% to 71.04% on VAT. The curves at tolerance 0.10 show the same overall positive association. These results show that higher predicted reliability identifies candidates with greater target proximity, consistent with its role as a source-weighting prior in Equation (5).

E.1 CANDIDATE SELECTION ACROSS POOL SIZES

We compared learned reliability with coarse-heatmap scores on candidate pools of $K \in \{8, 10, 100\}$. For each query and each K, both scoring rules received the same candidates selected by the peak-first rule in Equation (S4), with K replacing eight. The heatmap score is the spatial-softmax probability at each candidate on the 32×32 grid. We reused full-model features and evaluated the already trained reliability network with frozen weights. The evaluated network uses eight candidates; the larger pools are separate scoring probes that leave its final heatmap unchanged.

We evaluated all 4,782 GazeFollow and 20,655 VAT in-frame test queries. Each rule selected its highest-scoring candidate; exact ties were averaged over all maximizers. H_δ denotes Hit@1: the percentage of selected candidates within normalized distance δ of any GazeFollow annotation or the single VAT target. Average L_2 measures distance to the mean GazeFollow annotation or the VAT target, using the same endpoint-inclusive candidate coordinates for both rules.

Table 10: Coarse-heatmap and reliability ranking across all evaluated candidate-pool sizes. $H_{0.05}$ and $H_{0.10}$ are percentages. Bold red and underline mark the better and lower-ranked values within each two-row comparison at fixed K. These are candidate-selection metrics. Distances are measured at the selected candidate locations, rather than the final heatmap predictions reported in Tables 1 and 2.

		GazeFollow			VideoAttentionTarget		
K	Score	$H_{0.05}\uparrow$	$H_{0.10}\uparrow$	Avg $L_2\downarrow$	$H_{0.05}\uparrow$	$H_{0.10}\uparrow$	$L_2\downarrow$
8	Heatmap	**85.902**	93.862	0.089282	69.342	77.758	0.081154
8	Reliability	85.738	**93.904**	**0.089130**	**69.371**	**77.887**	**0.080706**
10	Heatmap	**85.902**	93.862	0.089282	**69.342**	77.758	0.081154
10	Reliability	85.822	**93.967**	**0.088798**	69.322	**77.887**	**0.080780**
100	Heatmap	**85.902**	93.862	0.089282	**69.342**	77.758	0.081154
100	Reliability	85.473	**94.152**	**0.088075**	69.060	**78.023**	**0.080394**

At $K = 100$, the pools contain an average of 13.29 local peaks on GazeFollow and 18.98 on VAT; the remaining positions enter through the non-maximal fallback rule. The larger pools therefore include additional locations around and beyond the selected peaks.

Relative to heatmap ranking at $K = 100$, reliability reduces average L_2 from 0.089282 to 0.088075 on GazeFollow and from 0.081154 to 0.080394 on VAT, relative reductions of 1.35% and 0.94%. $H_{0.10}$ increases from 93.862% to 94.152% and from 77.758% to 78.023%, respectively. $H_{0.05}$ decreases from 85.902% to 85.473% on GazeFollow and from 69.342% to 69.060% on VAT. In these candidate-selection probes, reliability improves average target distance and the broader localization tolerance.

Using 2,000 paired bootstrap resamples of GazeFollow images and VAT videos, the 95% intervals for the reliability-minus-heatmap average-distance difference at $K = 100$ are $[-0.001986, -0.000416]$ and $[-0.001665, 0.000034]$, respectively. These intervals quantify test-sample uncertainty for the fixed checkpoint.

E.2 CANDIDATE-SCORE VISUALIZATION

In these examples, reliability moves the top-ranked candidate from a competing region to an annotated target, reducing $d_{\min}$ from 0.555 to 0.013, from 0.543 to 0.012, and from 0.277 to 0.019. Table 10 reports the corresponding candidate-selection metrics over the complete test sets.

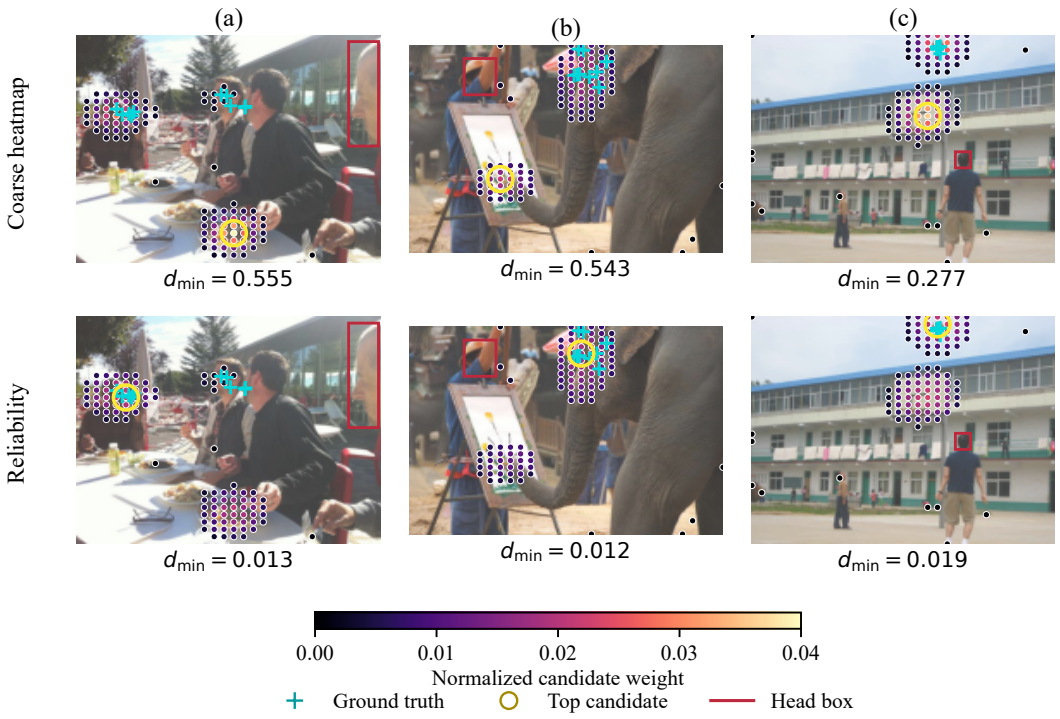

Figure 7: Candidate-score visualization on GazeFollow. Each column shows a coarse-heatmap ranking error corrected by reliability, with coarse-heatmap scores above and reliability scores below on the same 100 candidates. Scores are floored at 10^{-8} and normalized by their sum within each panel; all panels share a linear color scale. Cyan crosses denote gaze annotations, red boxes identify the queried head, and yellow rings mark the highest-scoring candidate. The normalized distance $d_{\min}$ is measured to the nearest gaze annotation.