MODELS ALREADY KNOW WHERE TO LOOK: RECOVERING THE DISCARDED SPATIAL MAP IN PARALLEL SCANPATH PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Task-driven scanpath prediction models an ordered, variable-length sequence of fixations made during a visual task. Some parallel models supervise a spatial distribution for each fixation slot but discard it at inference, using an independent coordinate-regression head instead. We study this representation–decoding mismatch through a controlled family of models. C0 uses coordinate regression, S0 adds an auxiliary spatial map, and S0-SpatialDecode (S0-SD) decodes the map of an already trained S0 checkpoint without retraining. Task-Conditioned Spatial Decoding (TCSD) instead trains and predicts through the same spatial distribution. On an image-disjoint COCO-Search18 evaluation, S0-SD improves all six metrics over S0 under joint seed–image resampling. A stronger regression head narrows but does not close the gap to spatial decoding. S0-SD and TCSD show no reliable difference on any metric, indicating that most of the gain comes from using the learned map at inference rather than from the TCSD training pathway. The effect persists with ImageNet-pretrained features, while experiments on Gazeformer and AiR-D show that its transfer is metric- and task-dependent. TCSD retains efficient one-pass inference, although HAT-SF remains more accurate on most metrics. These results show that spatial supervision and the decoder that turns the learned map into fixation coordinates should be evaluated together.

1 INTRODUCTION

Task-driven scanpath prediction seeks to model the ordered, variable-length sequence of fixations made by an observer while performing a visual task Yang et al. (2020). Unlike static saliency prediction, it must model not only where observers look, but also fixation order, duration, and termination Chen et al. (2021). Recent methods address this problem by learning task-conditioned visual features and decoding them into a sequence of fixation locations.

Parallel scanpath models predict several ordered fixation slots in a single forward pass Mondal et al. (2023). One possible design is to supervise a dense spatial distribution for each slot while relying on a separate regression head to produce the final coordinate. The spatial head may then learn a useful localization signal without directly affecting the location predicted at test time. This creates a representation–decoding mismatch: the model is trained to identify where a fixation should occur in one representation, but uses another representation to determine its output.

We examine whether the supervised spatial distribution contains localization information missed by the regression head, and whether recovering this information requires retraining or only a change in inference-time decoding. To isolate these effects, we construct a controlled family of parallel scanpath predictors. C0 uses direct coordinate regression. S0 adds a supervised, task-conditioned spatial distribution but retains the regression head as its location output. S0-SpatialDecode (S0-SD) uses the same trained S0 parameters and instead decodes coordinates from its spatial distribution, without retraining. Finally, Task-Conditioned Spatial Decoding (TCSD) trains and predicts through a single spatial representation.

In TCSD, each ordered fixation slot is matched against task-conditioned image features to produce a categorical distribution over spatial cells. The predicted coordinate is the expectation of the cell centers under this distribution. The same distribution therefore receives map-level supervision,

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.

contributes to coordinate-level supervision, and determines the final fixation location. Expectation decoding is a standard differentiable operation; TCSD uses it to remove the independent regression bypass and align the supervised representation with the inference-time output.

On an image-disjoint COCO-Search18 evaluation, TCSD achieves higher three-seed means than C0 and S0 across all six sequence- and saliency-based metrics. Paired physical-image bootstrap intervals, with the three trained seeds fixed, also favor TCSD over S0 on every metric. Decoding S0's existing spatial distribution improves all six metrics without changing its parameters or training procedure. S0-SD recovers most of the gain and attains slightly better three-seed means than TCSD on five metrics, while TCSD performs better on KLD. When both training seeds and physical images are resampled, however, the confidence intervals for all six differences between S0-SD and TCSD include zero.

We also test whether the result can be explained by an underpowered coordinate head. Replacing S0's linear coordinate output with a two-layer MLP supplied with pooled task-conditioned visual context improves S0 on five of the six metrics. Nevertheless, TCSD retains better three-seed means on all six metrics, and every paired physical-image bootstrap interval favors TCSD over this stronger regression control. These findings indicate that regression capacity matters, but does not recover all of the localization information available in the spatial distribution. Most of the improvement over S0 therefore comes from using the learned spatial map at test time, while the current three-seed evidence does not establish a separate advantage for training through TCSD rather than decoding the map post hoc.

We further compare TCSD with Gazeformer Mondal et al. (2023), ScanDiff Cartella et al. (2025), an IRL-based predictor Yang et al. (2020), and a shared-feature HAT adaptation Yang et al. (2024). HAT-SF remains more accurate on most metrics, while TCSD retains efficient one-pass parallel inference. Experiments with Gazeformer test whether the same decoding effect appears in another parallel architecture, and experiments on AiR-D Chen et al. (2020) examine its behavior under natural-language question conditioning.

Our contributions are:

- We identify a representation–decoding mismatch in parallel scanpath prediction: a supervised spatial map can retain useful localization information that is bypassed by the model's inference-time regression head.
- We show that decoding S0's already-trained spatial distribution improves all six evaluation metrics without retraining and recovers most of the improvement achieved by TCSD. A stronger MLP regression control improves S0 but does not close the gap to spatial decoding.
- We present TCSD as an end-to-end aligned implementation of spatial decoding and evaluate it using an image-disjoint, multi-seed protocol with paired bootstrap analyses, controlled ablations, external baselines, efficiency measurements, and tests on Gazeformer and AiR-D.

2 Related Work

From saliency maps to scanpaths. Saliency prediction estimates the aggregate likelihood that different image regions attract gaze Borji & Itti (2012); Jetley et al. (2016). Although a saliency map describes where observers tend to look, it does not specify the order of their fixations Wloka et al. (2018), how long each fixation lasts Chen et al. (2021), or when viewing terminates Chen et al. (2022). Scanpath prediction instead models gaze as an ordered, variable-length sequence. Depending on the setting, this sequence may be driven by image content alone or conditioned on a goal such as visual search or question answering.

Task-driven scanpath prediction. Task-driven methods use a target, question, or instruction to determine which image regions are relevant. Yang et al. Yang et al. (2020) formulate visual search as a sequential decision problem, while HAT Yang et al. (2024) predicts spatial fixation maps autoregressively. Gazeformer Mondal et al. (2023) replaces sequential generation with a set of ordered fixation queries that are decoded in parallel, and ScanDiff Cartella et al. (2025) uses a diffusion process to generate scanpaths. Other work broadens the task beyond location prediction: GazeX-

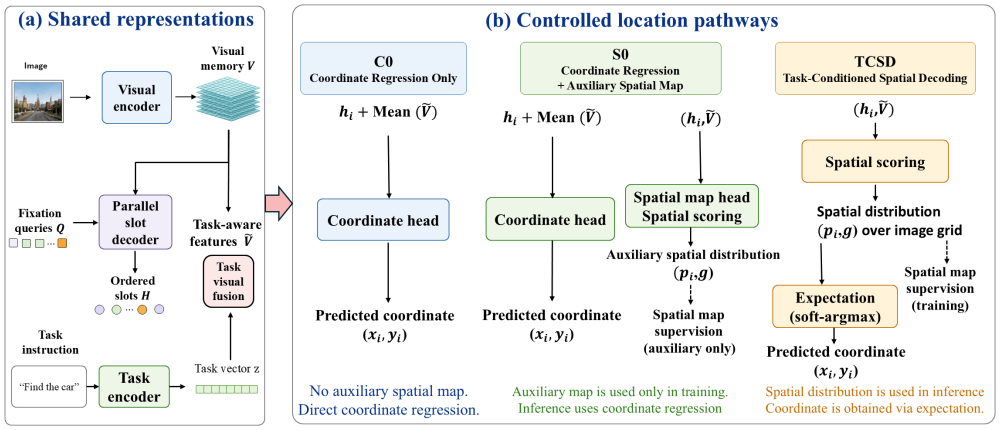

Figure 1: Controlled comparison of the core configurations. All models share the encoder, ordered fixation slots, duration head, and termination head. C0 uses coordinate regression; S0 adds a supervised spatial map but retains regression at inference; TCSD decodes the coordinate from that map. Dashed arrows denote training-only supervision.

plain Chen et al. (2024) generates fixation-level language explanations together with gaze, and AiR-D Chen et al. (2020) records question-conditioned eye movements during visual reasoning. These approaches differ not only in how they represent the task, but also in how a latent fixation state is converted into an image location.

Representing and decoding fixation locations. Existing scanpath models illustrate two common ways of obtaining fixation locations. A coordinate head produces an x–y point directly, as in Gazeformer Mondal et al. (2023). A spatial head instead predicts a distribution over image regions. GazeXplain, for example, samples fixation locations from step-wise spatial distributions Chen et al. (2024), whereas HAT generates fixations autoregressively from dense spatial maps Yang et al. (2024). FlowScan extends the latter framework with self-supervised visual features, soft-argmax decoding, and an auxiliary flow-matching loss used during training Aklilu et al. (2026). Spatial outputs can therefore serve as auxiliary supervision, as the source of the predicted coordinate, or both.

Extracting a coordinate by taking the expectation of a spatial distribution is a standard differentiable operation, widely used in heatmap-based localization Sun et al. (2018); Nibali et al. (2018). The unresolved issue in parallel scanpath prediction is not whether this operation is available, but which learned representation should determine the final fixation. When a supervised spatial map is paired with an independent regression head, improvements may come from the additional map supervision even though the map is discarded at test time. Alternatively, the map itself may contain localization information that the regression head fails to recover. Comparisons among these architectures do not isolate the two effects, because their training objectives, decoding procedures, and model capacity differ at the same time.

Our experiments isolate this distinction within a shared parallel architecture. S0 tests spatial-map supervision while retaining coordinate regression, S0-SpatialDecode uses the same trained model but decodes its auxiliary map at test time, and TCSD trains and predicts through that map end to end. This design separates the effect of learning a spatial distribution from the effect of using it to produce fixation coordinates.

3 Methodology

We study whether the representation receiving spatial supervision is also used to produce fixation coordinates. The models in Figure 1 share the same parallel architecture and differ only in their location pathway.

3.1 Problem and Shared Architecture

Given an image I and task condition c, the model predicts an ordered scanpath

$$\mathcal{S} = [(\mathbf{x}_i, d_i)]_{i=1}^{T}, \qquad \mathbf{x}_i = (x_i, y_i), \tag{1}$$

where d_i is fixation duration and T is determined by a learned termination head. Coordinates are normalized by image width and height to $\bar{\mathbf{x}}_i \in [0,1]^2$.

Let $\mathbf{F} \in \mathbb{R}^{G \times D_{\text{img}}}$ denote precomputed features on an $H_g \times W_g$ grid, where $G = H_g W_g$. A visual Transformer Vaswani et al. (2017) produces

$$\mathbf{V} = E_{\text{vis}}(\mathbf{F}) = [\mathbf{v}_1, \ldots, \mathbf{v}_G] \in \mathbb{R}^{G \times D}. \tag{2}$$

The task condition is encoded by a frozen RoBERTa Transformer Liu et al. (2019) with a trainable pooler and projection, producing $\mathbf{z} \in \mathbb{R}^D$. It is fused with each visual feature through a linear layer, ReLU, and dropout:

$$\widetilde{\mathbf{v}}_g = F_{\text{fuse}}([\mathbf{v}_g; \mathbf{z}]). \tag{3}$$

The model maintains L learned ordered queries $\mathbf{Q} = [\mathbf{q}_1, \ldots, \mathbf{q}_L]$. Following query-based Transformer decoding Carion et al. (2020), they attend to the visual memory and produce slot states

$$[\mathbf{h}_1, \ldots, \mathbf{h}_L] = D_{\text{slot}}(\mathbf{Q}, \mathbf{V}). \tag{4}$$

Slot i represents fixation i and predicts a location, duration, and termination probability. Sequences longer than L are truncated; unused slots are masked.

3.2 Location Pathways

Coordinate regression. C0 and S0 mean-pool the task-conditioned visual features, $\bar{\mathbf{v}}_c = G^{-1} \sum_g \widetilde{\mathbf{v}}_g$, and predict

$$\hat{\mathbf{x}}_i^{\text{reg}} = \text{sigmoid}\left(f_x(\mathbf{h}_i + \bar{\mathbf{v}}_c)\right). \tag{5}$$

C0 is supervised only through this coordinate output. S0 retains the same output but additionally learns the spatial distribution below.

Spatial distribution. For each slot, projected slot and image features are matched by a scaled dot product:

$$s_{i,g} = \frac{f_q(\mathbf{h}_i)^\top f_k(\widetilde{\mathbf{v}}_g)}{\sqrt{D}}, \qquad p_{i,g} = \frac{\exp(s_{i,g})}{\sum_{j=1}^{G} \exp(s_{i,j})}. \tag{6}$$

Both S0 and TCSD supervise p_i, but only TCSD uses it to produce the coordinate. If $\mathbf{r}_g$ is the normalized center of cell g, TCSD predicts

$$\hat{\mathbf{x}}_i^{\text{TCSD}} = \sum_{g=1}^{G} p_{i,g} \mathbf{r}_g. \tag{7}$$

This expectation is differentiable, so map and coordinate supervision update the same spatial scores. TCSD has no independent coordinate head.

Post-hoc spatial decoding. S0-SpatialDecode (S0-SD) applies Eq. equation 7 to the auxiliary map of an already trained S0 checkpoint. No parameters are retrained, and its map, duration, and termination predictions are unchanged. Comparing S0-SD with S0 therefore isolates the effect of inference-time coordinate decoding.

3.3 Regression and Decoding Controls

We use three controls to test alternative explanations. S0-MLP replaces S0's linear coordinate projection with a two-layer MLP operating on $[\mathbf{h}_i; \bar{\mathbf{v}}_c]$, while retaining the auxiliary map objective. TADR lets the slot decoder attend directly to $\widetilde{\mathbf{V}}$ and predicts coordinates without a spatial map. SMR retains TCSD's spatial matching but replaces expectation decoding with a learned regressor applied to

$$\left[\mathbf{h}_i; \sum_{g=1}^{G} p_{i,g} \widetilde{\mathbf{v}}_g\right]. \tag{8}$$

These controls separate coordinate-head capacity, task-aware decoder memory, spatial matching, and expectation decoding. Their full definitions are given in the appendix.

3.4 Training Objectives

Let m_i^{fix} indicate a valid fixation and $N_{\text{fix}} = \sum_i m_i^{\text{fix}}$. In the primary experiments, the coordinate target is the center $\mathbf{r}_{g_i^*}$ of the cell containing fixation i. The coordinate loss is

$$\mathcal{L}_{\text{coord}} = \frac{1}{N_{\text{fix}}} \sum_{i=1}^{L} m_i^{\text{fix}} \ell_{\text{SL1}} \left(\hat{\mathbf{x}}_i, \mathbf{r}_{g_i^*}\right). \tag{9}$$

An exact-coordinate target is examined as a control. For map supervision, q_i is a normalized Gaussian centered at the recorded fixation, with distance measured in grid cells. The map loss is

$$\mathcal{L}_{\text{map}} = -\frac{1}{N_{\text{fix}}} \sum_{i=1}^{L} m_i^{\text{fix}} \sum_{g=1}^{G} q_{i,g} \log p_{i,g}. \tag{10}$$

All configurations share a log-normal duration objective $\mathcal{L}_{\text{dur}}$ and a binary termination objective $\mathcal{L}_{\text{stop}}$. The complete loss is

$$\mathcal{L} = \mathcal{L}_{\text{stop}} + \lambda_d \mathcal{L}_{\text{dur}} + \alpha_m \mathcal{L}_{\text{map}} + \lambda_x \mathcal{L}_{\text{coord}}. \tag{11}$$

C0 and TADR use $(\alpha_m, \lambda_x) = (0, 1)$; S0, S0-MLP, TCSD, and SMR use $(1, 1)$. Coordinate-only TCSD-C and map-only TCSD-M use $(0, 1)$ and $(1, 0)$, respectively. All active weights are one. Details of the duration and termination losses are given in the appendix.

3.5 Inference

Each model uses its corresponding location pathway at inference. S0-SD is the sole exception: it replaces S0's regression output with Eq. equation 7. Duration is decoded as the median of the predicted log-normal distribution. Slots are read in order, and decoding stops before the first slot whose termination probability is at least 0.5; otherwise, all L fixations are retained. Coordinates are finally rescaled to image pixels and durations to milliseconds.

Model	MM ↑	SM ↑	SED ↓	KLD ↓	CC ↑	NSS ↑
C0	0.783 ± 0.005	0.247 ± 0.040	4.942 ± 1.567	19.835 ± 0.118	0.081 ± 0.005	0.430 ± 0.024
S0	0.816 ± 0.004	0.365 ± 0.032	3.789 ± 0.515	16.763 ± 1.352	0.132 ± 0.032	0.713 ± 0.197
S0-MLP	0.822 ± 0.002	0.391 ± 0.027	3.828 ± 1.107	14.399 ± 0.162	0.213 ± 0.026	1.271 ± 0.223
S0-SD†	$\mathbf{0.832} \pm 0.001$	$\mathbf{0.472} \pm 0.015$	$\mathbf{3.333} \pm 0.531$	11.362 ± 0.446	$\mathbf{0.332} \pm 0.023$	$\mathbf{2.107} \pm 0.184$
TCSD	0.831 ± 0.002	0.470 ± 0.012	3.545 ± 0.429	$\mathbf{11.104} \pm 0.146$	0.331 ± 0.011	2.041 ± 0.131

Table 1: COCO-Search18 results, reported as mean ± sample standard deviation over three seeds. MM and SM denote MultiMatch and ScanMatch. Bold denotes the best mean. †S0-SD reuses the S0 checkpoints and changes only inference-time coordinate decoding.

Metric	S0-SD vs. S0	TCSD vs. S0-MLP†	TCSD vs. S0-SD
MultiMatch	0.0160 [0.0107, 0.0215]	0.0084 [0.0061, 0.0108]	−0.0011 [−0.0041, 0.0018]
ScanMatch	0.1076 [0.0774, 0.1410]	0.0786 [0.0685, 0.0890]	−0.0026 [−0.0186, 0.0143]
SED	0.4554 [0.2977, 0.6083]	0.2837 [0.0544, 0.5071]	−0.2116 [−0.6687, 0.3026]
KLD	5.4011 [3.7797, 6.9479]	3.2953 [2.9379, 3.6709]	0.2579 [−0.2091, 0.7297]
CC	0.1995 [0.1496, 0.2425]	0.1176 [0.1019, 0.1333]	−0.0012 [−0.0301, 0.0232]
NSS	1.3935 [1.0450, 1.7019]	0.7700 [0.6457, 0.8975]	−0.0662 [−0.3188, 0.1386]

Table 2: Direction-normalized effects and 95% percentile intervals; positive values favor the first method. The first and third comparisons resample matched seeds and images. †The S0-MLP comparison resamples images while averaging the three fixed seeds.

4 Experiments

4.1 Experimental Setup

We use the target-present subset of COCO-Search18 under official split 1 Yang et al. (2020). The test set contains 3,258 scanpaths from 315 physical images. We remove 79 training scanpaths whose eight images overlap with the test set and reserve 10% of the remaining training images for development. The partitions are image-disjoint but not observer-disjoint. AiR-D Chen et al. (2020) is evaluated separately for question-conditioned gaze.

The shared model uses a 24×32 spatial grid, 16 fixation slots, two encoder and two decoder layers, hidden dimension 256, and eight attention heads. We train with AdamW for at most 30 epochs using early stopping. Unless stated otherwise, results are means and sample standard deviations over seeds 3, 4, and 5. Further implementation details are provided in the appendix.

We report MultiMatch, ScanMatch, SED, KLD, CC, and NSS. Higher is better except for SED and KLD. All same-protocol comparisons use the same test images, coordinate conventions, termination rule, and evaluator. Matched comparisons use 20,000 paired bootstrap replicates over physical images; the main intervals also resample the three training seeds. These intervals test robustness to the observed seeds rather than precisely estimating training randomness.

4.2 Controlled Comparison

Table 1 isolates the location pathway. S0 improves over coordinate-only C0 by adding spatial supervision, while TCSD improves over both on all six mean scores. S0-SD then changes only the inference-time decoder of the trained S0 checkpoints. It improves all six metrics without retraining and recovers most of the TCSD gain. S0-SD has better means on five metrics, while TCSD is better on KLD.

Table 2 gives the paired comparisons. S0-SD improves over S0 on every metric under joint seed–image resampling, with the same direction for each individual seed. By contrast, every TCSD–S0-SD interval contains zero. The evidence supports decoding the learned map at inference, but does not show that the TCSD training pathway is better than post-hoc spatial decoding.

S0-MLP shows that coordinate-head capacity matters, but it remains below TCSD on every mean, with all paired image intervals favoring TCSD. Further controls reach the same conclusion: task-aware regression improves over C0, whereas spatially matched regression remains below expectation



Figure 2: Qualitative COCO-Search18 example for the target *bottle*. From left to right: the human scanpath, S0's native regression output, S0-SD, and TCSD. Numbered markers and connecting lines indicate fixation order; the heatmap overlays show the spatial distributions used for decoding. S0-SD reuses the S0 checkpoint and changes only the coordinate source, moving later fixations toward the target bottle.

decoding. Coordinate-only and map-only TCSD variants are close to the full model on most metrics. Complete ablations are reported in the appendix.

4.3 Robustness and Scope

The decoding gain is not specific to COCO pretraining. With ImageNet-pretrained features, S0-SD again improves all six metrics, and a difference-in-differences analysis finds no reliable change from the COCO-pretrained setting.

Transfer to Gazeformer is more limited. Decoding its auxiliary map from the same checkpoints improves MultiMatch from 0.8153 to 0.8229 (95% interval $[0.0047, 0.0108]$); ScanMatch, SED, CC, and NSS are inconclusive, and KLD becomes worse. On AiR-D, S0-SD improves five of six means but not MultiMatch. Its predictions move only 8.6–9.0 pixels when the question changes, compared with a descriptive matched-slot displacement of 122.9 pixels between human trials. These quantities are not directly comparable, but they indicate that question sensitivity remains a limitation. Full transfer results are provided in the appendix.

4.4 Map Geometry and Qualitative Example

Across images, larger S0-SD gains are associated with lower map entropy, fewer spatial modes, and higher peak probability. These associations are correlational, but they suggest that expectation decoding is most reliable when the learned distribution is spatially concentrated. Full correlation estimates and intervals are reported in the appendix.

Figure 2 illustrates the representation–decoding mismatch. S0's regression output remains near the monitor and keyboard, even though its auxiliary distribution places probability on the target bottle. S0-SD changes only the coordinate source of the same checkpoint and redirects the later fixations toward the bottle, producing a path similar to TCSD. This example is illustrative; the quantitative comparisons in Tables 1 and 2 provide the aggregate evidence.

4.5 External Baselines and Efficiency

We retrain Gazeformer Mondal et al. (2023), ScanDiff Cartella et al. (2025), an IRL-based predictor Yang et al. (2020), and a shared-feature HAT adaptation (HAT-SF) Yang et al. (2024) under the same image-disjoint protocol and evaluator. TCSD is competitive on MultiMatch, but HAT-SF is better on five metrics. Their MultiMatch difference is inconclusive under paired image resampling (effect -0.0010, 95% interval $[-0.0048, 0.0028]$).

Published results are listed separately in the appendix because the source papers use different splits, horizons, sampling procedures, and metric definitions. The available official Gazeformer cluster artifact covers none of the frozen test task–image pairs, so we do not refit clusters using held-out labels.

The metric-dependent HAT-SF gap reflects different aspects of scanpath quality. MultiMatch emphasizes relative path geometry, whereas CC and NSS measure aggregate fixation density. Sequence length also affects SED: length normalization reverses the ordering of TCSD and Gazeformer (0.545

Model	MultiMatch ↑	ScanMatch ↑	SED ↓	KLD ↓	CC ↑	NSS ↑
Gazeformer	0.8114	0.4157	2.5974	12.9802	0.3058	2.1420
ScanDiff	0.7772	0.3641	2.9608	**6.9034**	0.2964	1.6388
IRL	0.8025	0.4090	3.0125	8.1645	0.4599	3.2597
HAT-SF	**0.8316**	**0.5151**	**2.1157**	9.3962	**0.5300**	**3.6814**
TCSD	0.8306	0.4699	3.5447	11.1039	0.3305	2.0407

Table 3: Three-seed means under the common image-disjoint protocol and evaluator. Bold denotes the best mean for each metric.

versus 0.586), although HAT-SF remains better at 0.471. The remaining gap cannot be attributed specifically to autoregressive decoding because the models also differ in capacity and optimization.

TCSD retains an efficiency advantage. On an RTX 4060 Laptop GPU, it requires 15.86 ms at batch size 1 and processes 320.65 images/s at batch size 8. It is $2.53\times$ faster than HAT-SF at batch size 1 and has $4.21\times$ higher batch-8 throughput. Full parameter, latency, and throughput measurements are reported in the appendix.

5 Discussion

The main result is not that spatial supervision alone improves scanpath prediction, but that its value depends on whether the learned representation is used to produce the final coordinate. S0-SD changes no trained parameters yet consistently improves over S0, while its difference from TCSD is not robust across seeds and images. This distinguishes representation learning from representation use: a model may learn a useful spatial map but fail to exploit it when an independent regression head determines the prediction. TCSD provides an end-to-end implementation of the aligned design, although the experiments do not establish an advantage for its training pathway over post-hoc spatial decoding.

The comparison suggests a simple diagnostic for models with auxiliary structured outputs. The native prediction can be compared with a deterministic decoding of the auxiliary representation from the same checkpoint. Because the parameters and upstream features remain fixed, the comparison isolates whether the representation is useful when it is actually used. If post-hoc decoding recovers the performance of an end-to-end aligned model, the improvement should not be attributed solely to its training objective. Although we test this diagnostic only for scanpath prediction, the same issue can arise whenever a supervised heatmap or distribution is paired with an independent regression head.

The robustness tests define the scope of the result. The effect persists with ImageNet rather than COCO pretraining, suggesting that it is not an artifact of COCO-derived visual features. Its transfer to Gazeformer is weaker and metric-dependent, however, and AiR-D shows that spatial decoding does not by itself produce strong sensitivity to natural-language questions. The representation–decoding mismatch is therefore a useful diagnostic principle rather than a guarantee that spatial decoding will improve every architecture, metric, or task.

Map geometry may help explain this variation. Larger decoding gains are associated with sharper, less multimodal distributions, whereas diffuse maps can place their expectation between plausible fixation regions. Decoders that preserve multiple spatial hypotheses may be preferable in such cases. The external comparisons also reveal an accuracy–efficiency trade-off: TCSD preserves one-pass parallel inference, whereas HAT-SF remains stronger on most accuracy metrics. The present evidence cannot attribute this gap specifically to parallel versus autoregressive generation because the models also differ in capacity and optimization.

6 Limitations

The main variants are trained with three random seeds. The joint seed–image bootstrap tests whether the conclusions survive the observed seed variation, but it cannot precisely characterize training randomness. Evidence beyond the base architecture is also limited to one additional model, Gaze-

former, and one additional dataset and task formulation, AiR-D; the Gazeformer result is not consistent across all metrics. Our comparison with HAT-SF does not control capacity, optimization, and compute simultaneously, so it cannot isolate the effect of parallel versus autoregressive decoding. Expectation decoding also has an inherent weakness for multimodal maps, since the mean may lie between plausible fixation regions. On AiR-D, the human and model question-sensitivity measurements come from different experimental conditions and should therefore be interpreted only as a descriptive comparison.

7 Conclusion

This paper asks whether a supervised spatial map is useful when it is actually used to predict fixation locations. Decoding the map already learned by S0 improves all six COCO-Search18 metrics without retraining and recovers most of the gain achieved by TCSD. TCSD offers an end-to-end design in which supervision and prediction share the same spatial representation, but we find no robust evidence that its training pathway is better than post-hoc decoding of the S0 map. The practical lesson is that spatial supervision and spatial decoding should be assessed together: learning a useful map is not enough if the model bypasses it at inference.

AI use statement

In this work, we used generative AI tools to provide feedback on experimental design and manuscript presentation, and to assist with code editing, debugging, and optimization. We also used generative AI tools to edit and polish the manuscript for clarity, grammar, and readability. Generative AI tools were not used to independently generate the core research claims or independently interpret the experimental results.

All AI-assisted suggestions were critically reviewed by the authors. Experimental design suggestions were independently assessed before implementation, manuscript revisions were manually checked for accuracy and consistency with the underlying research, and AI-assisted code modifications were inspected and tested before being used in the experiments. The authors take full responsibility for the final content of this work, including all text, claims, experimental results, and software artifacts.

Ethics statement

We collect no new human data, use both datasets under their release terms, and do not redistribute images or gaze annotations. The AiR code is MIT-licensed, but its annotations have no separate stated license. Upon acceptance, we will release our code and evaluator under MIT and checkpoints for non-commercial research, subject to the data terms. Monitoring or assessment uses require consent and further validation.

References

Brahan Aklilu, Ofir Itzhak Shahar, and Ohad Ben-Shahar. Flowscan: Self-supervised features and flow matching regularization for gaze scanpath prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5822–5831, 2026.

Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):185–207, 2012.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16206–16216, 2025.

Shi Chen, Ming Jiang, Jinhui Yang, and Qi Zhao. Air: Attention with reasoning capability. In *European Conference on Computer Vision*, pp. 91–107. Springer, 2020.

Xianyu Chen, Ming Jiang, and Qi Zhao. Predicting human scanpaths in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10876–10885, 2021.

Xianyu Chen, Ming Jiang, and Qi Zhao. Gazexplain: Learning to predict natural language explanations of visual scanpaths. In *European Conference on Computer Vision*, pp. 314–333. Springer, 2024.

Yupei Chen, Zhibo Yang, Souradeep Chakraborty, Sounak Mondal, Seoyoung Ahn, Dimitris Samaras, Minh Hoai, and Gregory Zelinsky. Characterizing target-absent human attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5031–5040, 2022.

Saumya Jetley, Naila Murray, and Eleonora Vig. End-to-end saliency mapping via probability distribution prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5753–5761, 2016.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Sounak Mondal, Zhibo Yang, Seoyoung Ahn, Dimitris Samaras, Gregory Zelinsky, and Minh Hoai. Gazeformer: Scalable, effective and fast prediction of goal-directed human attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1441–1450, 2023.

Aiden Nibali, Zhen He, Stuart Morgan, and Luke Prendergast. Numerical coordinate regression with convolutional neural networks. *arXiv preprint arXiv:1801.07372*, 2018.

Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 529–545, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Calden Wloka, Iuliia Kotseruba, and John K Tsotsos. Active fixation control to predict saccade sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3184–3193, 2018.

Zhibo Yang, Lihan Huang, Yupei Chen, Zijun Wei, Seoyoung Ahn, Gregory Zelinsky, Dimitris Samaras, and Minh Hoai. Predicting goal-directed human attention using inverse reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 193–202, 2020.

Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Ruoyu Xue, Gregory Zelinsky, Minh Hoai, and Dimitris Samaras. Unifying top-down and bottom-up scanpath prediction using transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1683–1693, 2024.