FlowScan: Self-Supervised Features and Flow Matching Regularization for Gaze Scanpath Prediction

Brahan Aklilu* Ofir Itzhak Shahar* Ohad Ben-Shahar
Stein Faculty of Computer and Information Science
Ben-Gurion University of the Negev, Israel
brahan@post.bgu.ac.il, shofir@post.bgu.ac.il, ben-shahar@cs.bgu.ac.il

Abstract

Predicting human scanpaths during visual search is important for understanding goal-directed attention and for building attention-aware systems. Recent work, notably HAT [30], has established strong baselines using transformer architectures with supervised ResNet-50 backbones. We propose FlowScan, which modifies the HAT framework in three ways: (1) it replaces the supervised backbone with a frozen self-supervised DINOv3 [26] vision transformer; (2) it adapts the deformable pixel decoder to operate on frozen ViT tokens and introduces soft-argmax fixation decoding; and (3) it adds a flow matching (FM) [19] auxiliary head that regularizes the decoder's internal representation, even though this head is discarded at inference. We evaluate FlowScan on the COCO-Search18 dataset [4] under the target-present condition. FlowScan achieves a Sequence Score of 0.603 *and a Semantic Sequence Score of* 0.538*, outperforming HAT retrained on the same test split (SS* 0.575*, SemSS* 0.499*) across all five evaluation metrics. Ablation studies show that each component contributes meaningfully: self-supervised ViT features improve spatial similarity over ResNet-50 (*$+12.4\%$ *SS), flow matching regularization yields a further* $+12.2\%$ *SS despite being discarded at inference, and the deformable pixel decoder adds another* $+3.8\%$ *SS through multi-scale feature fusion.*

1. Introduction

Visual search is a fundamental component of everyday human behavior. When looking for an object in a scene, the human visual system generates a sequence of fixations—a *scanpath*—that reflects the interplay of bottom-up saliency, top-down task guidance, and memory for previously inspected locations [12, 28]. Computationally predicting these scanpaths has applications in human-computer interaction, assistive technology, image retrieval, and in probing the cognitive mechanisms underlying visual attention [2].

Traditional saliency models predict *where* people look on average, producing a single probability map over locations [13, 16]. However, scanpath prediction is inherently sequential: each fixation depends on the task, the scene, and potentially on the full history of prior fixations. Recent deep learning approaches have therefore formulated scanpath prediction as an autoregressive generation task conditioned on the search target category [17, 23, 29, 30]. Among these, HAT (Human Attention Transformer) [30] currently represents the state of the art on the COCO-Search18 benchmark [4]. HAT uses a supervised ResNet-50 backbone [10] trained end-to-end, a multi-scale deformable attention pixel decoder, a foveated working memory, and a transformer encoder-decoder that autoregressively generates fixation heatmaps conditioned on a learned task query.

Despite its strong performance, HAT's backbone relies on supervised ImageNet features, which may not capture the rich semantic structure needed for task-guided visual search. Moreover, the model is trained purely with a heatmap-based loss, which supervises the peak of the predicted fixation distribution but only indirectly constrains the internal representation of fixation coordinates. This raises two questions: (i) can modern self-supervised ViT backbones provide stronger feature representations for scanpath prediction, and (ii) can we more directly regularize the decoder representation to encode precise fixation coordinates without increasing inference cost?

In this work, we propose FlowScan, which modifies the HAT architecture with three complementary improvements. **(1) Self-supervised backbone.** We replace HAT's supervised ResNet-50 with a frozen DINOv3 [26] vision transformer trained on a large-scale curated dataset (LVD-142M). Self-supervised features from DINO-family models encode rich semantic correspondences [3, 25], spatial structure, and object-level groupings without requiring labeled data. These properties are particularly well suited to visual search, where the model must understand scene semantics,

*Equal contribution.

object boundaries, and task-relevant regions.

(2) Adapted deformable pixel decoder with soft-argmax decoding. We adapt HAT's multi-scale deformable attention pixel decoder [31] to operate on frozen ViT patch tokens rather than end-to-end ResNet features, and reimplement deformable attention entirely in PyTorch without custom CUDA kernels. We also add soft-argmax fixation decoding to obtain sub-pixel precision from the predicted heatmaps.

(3) Flow matching regularization. We add an auxiliary flow matching [19] head that learns a velocity field transporting Gaussian noise to fixation coordinates, conditioned on the decoder's output. This head is trained jointly with the primary heatmap objective but is *not* used at inference. By providing a complementary gradient signal that directly supervises fixation coordinates, the flow matching head acts as a regularizer that improves the quality of the learned decoder representation without adding inference-time cost.

We evaluate FlowScan on the COCO-Search18 dataset under the target-present condition using standard scanpath metrics: Sequence Score (SS), Semantic Sequence Score (SemSS), and conditional saliency metrics (cNSS, cAUC, cIG). Our contributions are:

- We systematically demonstrate that frozen self-supervised ViT backbones substantially outperform a supervised ResNet-50 backbone for scanpath prediction ($+12.4\%$ SS).
- We show that flow matching regularization yields a $+12.2\%$ SS improvement despite being discarded at inference, acting as a powerful decoder representation regularizer.
- Using the authors' code, we retrain HAT on our data split for a fair head-to-head comparison and find that FlowScan outperforms HAT on all five metrics (SS 0.603 vs. 0.575, SemSS 0.538 vs. 0.499) under an identical evaluation protocol.

2. Related Work

2.1. Scanpath Prediction

Early scanpath prediction models were predominantly rule-based or relied on handcrafted features, often building on saliency-based attention mechanisms [2, 13]. Itti et al. [13] introduced an influential bottom-up saliency model based on feature contrast, later extended with inhibition-of-return to generate sequential fixations. Guided Search theory [27, 28] formalized the role of top-down task demands in visual search, motivating models that combine bottom-up saliency with task-conditioned guidance.

The COCO-Search18 dataset [4] enabled large-scale, data-driven scanpath prediction for goal-directed visual search by providing fixation sequences across 18 target categories. Yang et al. [29] proposed an inverse reinforcement learning (IRL) model that learns reward functions from human scanpaths for target-present search on COCO-Search18. Chen et al. [5] extended this work to target-absent search and introduced Fixation-to-Fixation Models (FFMs), which use recurrent networks to predict transitions between successive fixations. Gazeformer [23] adopted a transformer encoder-decoder architecture for scanpath prediction, demonstrating the benefit of attention mechanisms for modeling fixation sequences in goal-directed search.

HAT [30] unified various scanpath prediction tasks within a single transformer-based architecture. It introduced a foveated working memory that maintains both peripheral and foveal representations of the scene, processed by a transformer encoder-decoder with task-conditioned queries. HAT established the current state of the art on COCO-Search18 across multiple scanpath metrics, including sequence-similarity measures such as SemSS and MultiMatch [5, 8, 14].

Beyond HAT and Gazeformer, several recent models further explore scanpath prediction, including gaze scanpath transformers with spatiotemporal embeddings on COCO-Search18 [24] and object-level attention transformers that operate on sequences of object fixations on other visual search datasets [9].

More recently, UniAR [17] proposed a unified multi-task model that jointly predicts saliency maps, scanpaths, and subjective quality ratings across diverse visual domains using text-conditioned prompting. UniAR reported SemSS and MultiMatch scores on COCO-Search18 [5, 8, 14]; however, it uses a different evaluation protocol (no explicit target-present/target-absent split and a different metric subset) that prevents direct numerical comparison with HAT or our work.

2.2. Self-Supervised Visual Features

Self-supervised learning has produced visual representations that rival or exceed supervised features on many downstream tasks [3, 11, 25]. DINO [3] showed that self-supervised vision transformers learn features that encode semantic object boundaries and can support unsupervised segmentation. DINOv2 [25] scaled this approach with a larger curated dataset and improved training, yielding versatile features used across classification, segmentation, and depth estimation. DINOv3 [26] further refines this line of work, combining large-scale data curation with architectural improvements such as register tokens [7] to produce high-quality dense features for a wide range of vision tasks.

In the context of gaze prediction, backbone features directly influence the model's ability to represent task-relevant semantics. Kümmerer et al. [16] demonstrated that deeper supervised CNN features substantially improve fixation density prediction. To our knowledge, the potential of modern self-supervised features—such as DI-

NOv2/DINOv3 backbones—for scanpath prediction has not been systematically studied.

2.3. Flow Matching

Flow matching [19] is a simulation-free framework for training continuous normalizing flows. Given paired samples (x_0, x_1) from noise and data distributions, flow matching learns a velocity field $v_\theta(x_t, t)$ that transports x_0 to x_1 along the path $x_t = (1-t)x_0 + tx_1$. The training objective is:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t,x_0,x_1}\left[\|v_\theta(x_t, t \mid c) - (x_1 - x_0)\|^2\right], \quad (1)$$

where c is a conditioning context. In this work, we use flow matching not as a generative mechanism but as an *auxiliary regularizer* that encourages the model's decoder to encode precise fixation coordinates in its output representation.

3. Method

FlowScan builds on the HAT [30] framework with three modifications: (1) a frozen self-supervised backbone, (2) an adapted deformable pixel decoder with soft-argmax fixation decoding, and (3) a flow matching auxiliary head. We describe each component below.

3.1. Backbone

HAT uses a supervised ResNet-50 [10] backbone that is trained end-to-end alongside the rest of the model. We instead use a *frozen* self-supervised vision transformer, DINOv3 ViT-B/16 [7, 26], trained with self-distillation on a large-scale curated web dataset building on the LVD-142M corpus introduced in DINOv2 [25].

DINOv3 introduces register tokens [7] that reduce attention artifacts and improve patch feature quality. Input images are resized to 320×512, producing a 20×32 grid of 768-dimensional patch tokens. By freezing the backbone (85.7M parameters), only $\sim$11M parameters are trainable, enabling efficient single-GPU training in $\sim$4 hours.

3.2. Deformable Pixel Decoder

HAT's pixel decoder uses multi-scale deformable attention (MSDeformAttn [31]) over ResNet feature maps trained end-to-end. We adapt this design to operate on frozen ViT patch tokens and reimplement deformable attention in pure PyTorch, avoiding custom CUDA kernels. Backbone tokens are first projected to dimension $d = 256$ and reshaped into a 2D feature map, which is then downsampled via strided convolutions to form a 3-level feature pyramid network (FPN). Six layers of deformable attention (8 heads, 4 sampling points per head per level) process the multi-scale features, enabling cross-scale information exchange.

The decoder output is split into two feature levels:

- **P1** (coarse): a linear projection of the encoder output at the original grid resolution (20×32), used for peripheral scene context;
- **P4** (fine): a learnable $4\times$ upsampling via transposed convolution, producing a 79×127 feature map used for foveal crops and dense heatmap prediction.

Both levels have dimensionality $d = 256$.

3.3. Foveated Working Memory

Following HAT, we represent the visual input through a foveated working memory that mirrors the human visual system's dual-resolution processing. At each generation step, the memory consists of:

- **Peripheral tokens** ($7 \times 7 = 49$ tokens): obtained by adaptive average pooling of P1 to a fixed grid, providing coarse whole-scene context shared across all fixation steps.
- **Foveal tokens** ($7 \times 7 = 49$ tokens per fixation): cropped from P4 at each fixation location, capturing fine-grained spatial detail around the currently attended region.

Each token receives learnable scale embeddings (peripheral vs. foveal), temporal embeddings (fixation order), and spatial position embeddings. For a history of k fixations, the memory therefore contains $49 + 49k$ tokens in total.

3.4. Transformer Encoder-Decoder

The working-memory tokens are processed by a 3-layer transformer encoder with pre-norm layer normalization, 4 attention heads, and feedforward dimension $4d = 1024$. A task-conditioned decoder then cross-attends to the encoded memory.

Task conditioning. Following HAT, we use a learned embedding for each of the 18 COCO-Search18 target categories. The task embedding serves as the initial query to a 6-layer transformer decoder that cross-attends to the encoded memory. The decoder output is a single d-dimensional feature vector that encodes the predicted next fixation.

3.5. Output Heads

Dense heatmap head. The decoded query is projected through a 2-layer MLP ($d \to d$, ReLU) and compared against P4 features via a pixel-wise dot product. This produces an initial fixation score map, which is normalized with a spatial softmax to obtain a fixation probability distribution. A lightweight refinement CNN (Conv $1 \to 16$, GroupNorm, ReLU, Conv $16 \to 1$) then outputs the final fixation probability heatmap at P4 resolution (79×127). During training, this heatmap is supervised with a focal loss [18] against a Gaussian-blurred ground-truth fixation target ($\sigma = 1.5$ pixels; see Section 4.7).

Soft-argmax decoding. Rather than taking the raw argmax of the heatmap, we apply soft-argmax within a local 9×9

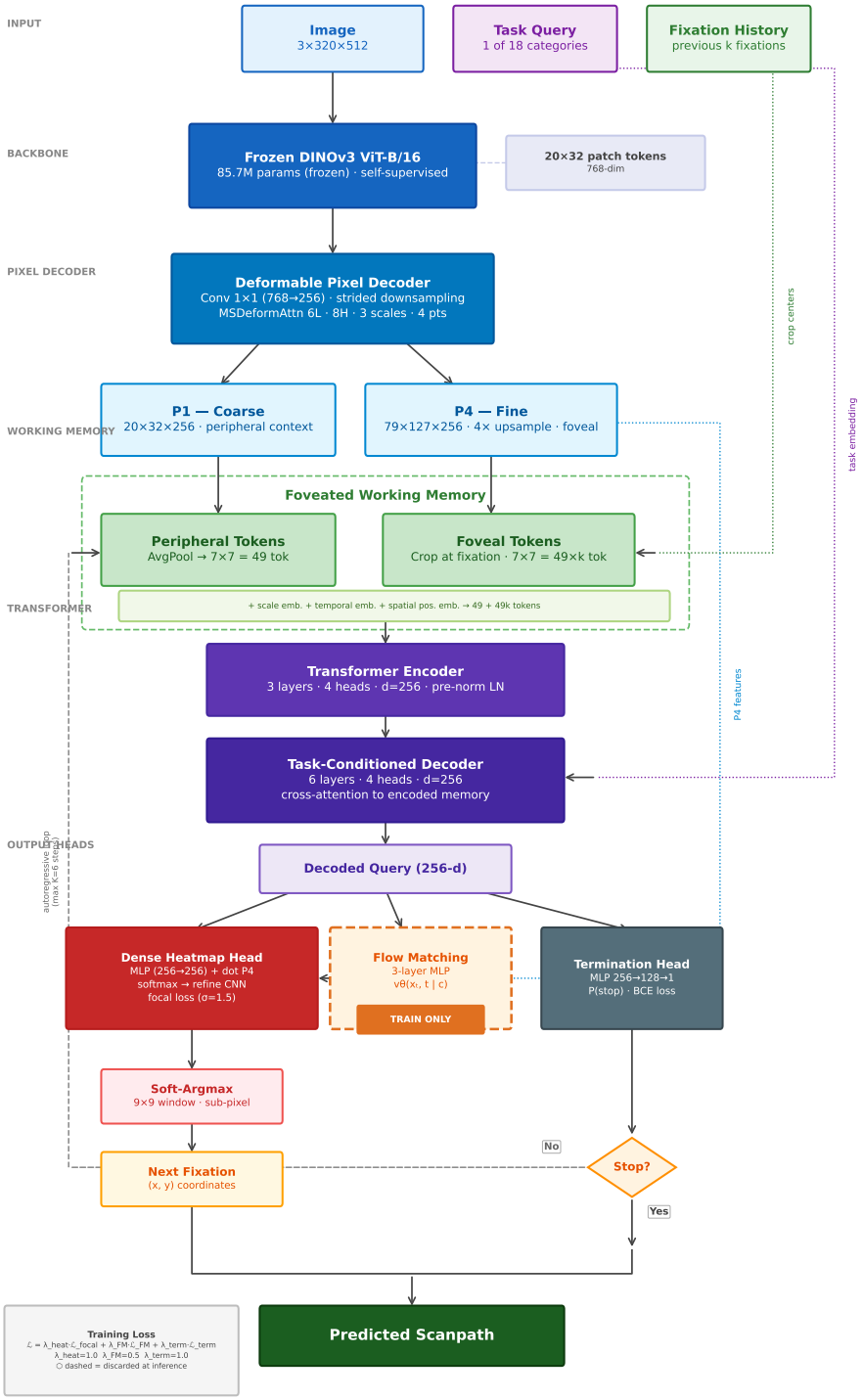


Figure 1. **FlowScan architecture.** A frozen DINOv3 backbone extracts patch tokens, which are processed by a deformable pixel decoder into coarse (P1) and fine-grained (P4) representations. Foveated Working Memory constructs a token sequence from peripheral context (pooled P1) and foveal crops (from P4 at each fixation location). A transformer encoder processes the memory, and a task-conditioned decoder generates a query that drives the dense heatmap head with soft-argmax decoding. The flow matching head (dashed) is trained as an auxiliary objective and discarded at inference.

window around its peak. The predicted fixation coordinate is the probability-weighted centroid within this window, providing sub-pixel precision.

Termination head. A two-layer MLP ($256 \to 128 \to 1$, ReLU, dropout) predicts the probability that the current fixation is the last in the scanpath, trained with binary cross-entropy.

Flow matching head (auxiliary). When enabled, a 3-layer MLP implements a velocity field $v_\theta(x_t, t \mid c)$ conditioned on the decoder output c. Given a ground-truth next fixation $x_1 \in [0,1]^2$ (normalized image coordinates) and noise $x_0 \sim \mathcal{N}(0, I)$, we construct the interpolant $x_t = (1-t)x_0 + tx_1$ for $t \sim \mathcal{U}(0,1)$ and regress the target velocity $v^* = x_1 - x_0$:

$$\mathcal{L}_{\text{FM}} = \|v_\theta(x_t, t \mid c) - v^*\|^2. \tag{2}$$

This auxiliary loss directly supervises fixation coordinates through the decoder features and complements the heatmap-based focal loss. The FM head contains only $\sim$150K parameters ($\sim$1.4% of the 11M trainable parameters), and its training overhead is negligible: a single forward–backward pass through a 3-layer MLP, with gradients flowing into the already-computed decoder features. Crucially, the flow-matching head is used only during training and is discarded at inference, so it introduces no additional cost during scanpath generation. We note that the learned velocity field could in principle be used at inference to generate diverse fixation samples via stochastic ODE integration, which we leave for future work (Section 5).

3.6. Training Objective

The total training loss $\mathcal{L}$ is defined as a weighted combination of three components:

$$\mathcal{L} = \lambda_{\text{heat}}\mathcal{L}_{\text{focal}} + \lambda_{\text{FM}}\mathcal{L}_{\text{FM}} + \lambda_{\text{term}}\mathcal{L}_{\text{term}}, \tag{3}$$

where $\mathcal{L}_{\text{focal}}$ and $\mathcal{L}_{\text{term}}$ are the primary objectives adopted from the HAT framework [30]. $\mathcal{L}_{\text{focal}}$ is the focal loss applied to the predicted fixation heatmap against a Gaussian-blurred ground truth, while $\mathcal{L}_{\text{term}}$ is a binary cross-entropy loss used to train the termination head. Our novel contribution, $\mathcal{L}_{\text{FM}}$, is the auxiliary flow matching loss defined in Eq. 2, which regularizes the decoder's internal representation by regressing target velocities toward fixation coordinates.

3.7. Scanpath Generation

At inference time, scanpaths are generated autoregressively. Given an image and task category, the model starts from an initial fixation at the image center (following the COCO-Search18 protocol [4]). At each step, the model: (1) builds the working memory from the current fixation history; (2) encodes the memory and decodes a task-conditioned query; (3) predicts a heatmap and selects the next fixation via soft-argmax; and (4) predicts a termination probability. Generation continues for up to $K = 6$ steps (excluding the initial fixation) or until the termination probability exceeds 0.5, following the HAT [30] protocol for target-present search.

4. Experiments

4.1. Dataset

We evaluate on COCO-Search18 [4], the standard benchmark for task-driven scanpath prediction. The dataset contains fixation sequences from 10 participants searching for 18 COCO object categories across 6,202 images. We focus on the **target-present (TP)** condition, where the search target exists in the image and scanpaths end when the target is fixated.

Data splits. As Coco-search's original testset is not publicly available, we randomly split the validation set to validation/test sets using a 70/30% ratio (seed 42), yielding 977 test scanpaths covering 319 image-task pairs. Since the exact test split used by HAT [30] was not publicly released, we retrain HAT on our split using the authors' code for a fair comparison. All models (ours and HAT retrained) use the same test split, ensuring valid head-to-head comparisons.

4.2. Evaluation Metrics

We report the following standard metrics, employed also by HAT [30]:

- **Sequence Score (SS):** Measures spatial similarity between predicted and ground-truth fixation sequences using MeanShift clustering [6] and Needleman-Wunsch string matching. Higher is better.
- **Semantic Sequence Score (SemSS):** Evaluates semantic similarity by mapping fixations to their underlying object categories using COCO panoptic segmentation [15] annotations (133 categories) and applying string matching. Higher is better.
- **Conditional Information Gain (cIG):** Information gain of predicted fixation density maps relative to a center-bias baseline, evaluated at ground-truth fixation locations [16]. Higher is better.
- **Conditional NSS (cNSS):** Normalized scanpath saliency of the predicted fixation density map, evaluated at ground-truth fixation locations. Higher is better.
- **Conditional AUC (cAUC):** Area under the ROC curve for the predicted fixation density map. Higher is better.

Following HAT, conditional metrics (cIG, cNSS, cAUC) are computed per-step, conditioning on the fixation history up to that point.

4.3. Implementation Details

All models use hidden dimension $d = 256$, 4 attention heads, 3 encoder layers, 6 decoder layers, and dropout 0.1.

Table 1. **Main results on COCO-Search18 target-present search.** ↑: higher is better. **Bold**: best, underline: second best. HAT retrained and FlowScan use the same test split (seed 42). HAT (published) uses a different split and is shown for reference only. SemSS is panoptic (80 "things" + 53 "stuff" = 133 COCO categories in total).

Method	SS↑	SemSS↑	cIG↑	cNSS↑	cAUC↑
Reference (different split, from [30]):					
IRL [29]	0.422	0.486	−9.709	1.977	0.913
Chen et al. [4]	0.445	0.518	−1.273	2.606	0.956
FFMs [5]	0.451	0.500	1.548	2.376	0.932
Gazeformer [23]	0.489	0.499	—	—	—
HAT [30]	0.470	0.543	2.399	5.086	0.977
Same test split:					
HAT retrained	0.575	0.499	2.491	4.649	0.915
FlowScan (ours)	**0.603**	**0.538**	**2.759**	**4.810**	**0.928**

The deformable pixel decoder uses 6 layers, 8 heads, 3 feature levels, and 4 sampling points. The foveated memory uses 7×7 (49) peripheral and foveal tokens. We train with AdamW [20] (learning rate 10^{-4}, weight decay 10^{-2}) with early stopping on validation loss. Regarding the loss defined in Eq. 3, we use $\lambda_{\text{heat}} = 1.0$, $\lambda_{\text{FM}} = 0.5$, and $\lambda_{\text{term}} = 1.0$. The selection of loss weights and the ground-truth heatmap scale (σ) are justified through empirical sensitivity analyses, detailed in Sec. 4.7. When flow matching is disabled (baseline models), $\mathcal{L}_{\text{FM}}$ is set to 0. The final model (deformable pixel decoder) trains for 20 epochs with batch size 24 and patience 7; FPN-based ablations use 30 epochs, batch size 32, and patience 10. Input images are resized to 320×512 (matching HAT). The backbone is frozen; only the pixel decoder, working memory, transformer encoder-decoder, and output heads are trained (∼11M trainable parameters). Training uses a single NVIDIA RTX 4090 GPU and takes approximately 4 hours per model. Ground-truth heatmaps use Gaussian blur with $\sigma = 1.5$ pixels (see Section 4.7).

For the HAT retrained baseline, we use the official HAT codebase with default hyperparameters on our data split.

4.4. Main Results

Table 1 presents our main comparison. All models in the "same test split" block are evaluated on identical data.

Published vs. retrained HAT. The original HAT results [30] were reported on a private data split and an evaluation protocol that are not publicly available. As shown in Table 1, retraining HAT on our open split yields a substantially higher Sequence Score (0.575 vs. 0.470) but lower values in other conditional metrics. These non-negligible shifts across all measures confirm that performance may be highly sensitive to the specific train/test partition and metric implementation. To ensure a mathematically rigorous and fair head-to-head comparison, we rely exclusively on the retrained HAT baseline. This ensures that both FlowScan and HAT are evaluated on identical images, using the exact same ground-truth scanpaths and the same code for metric calculation. HAT retrained on our split achieves substantially higher SS (0.575) than the published result (0.470), likely reflecting differences in data splits and evaluation protocol. We therefore rely exclusively on the retrained baseline for fair comparison.

FlowScan vs. HAT retrained. On our test split, FlowScan achieves SS = 0.603, outperforming HAT retrained (SS = 0.575) by +4.9% relative (+0.028 absolute). FlowScan also improves on all other metrics: SemSS (0.538 vs. 0.499, +7.8%), cIG (2.759 vs. 2.491), cNSS (4.810 vs. 4.649), and cAUC (0.928 vs. 0.915). Importantly, this comparison uses the same test split, same evaluation code, and same panoptic SemSS computation, providing a fair head-to-head evaluation.

Scanpath length. HAT retrained produces significantly shorter scanpaths than the ground truth (mean 2.58 vs. GT mean 3.68), with 60.5% of predictions being only 2 fixations long. FlowScan's mean prediction length (3.77) closely matches the ground truth (3.68), suggesting better-calibrated termination prediction.

Table 2. **Backbone comparison** (FPN pixel decoder, FM enabled). All backbones frozen. Same test split.

Backbone	Pre.	SS↑	SemSS↑	cIG↑	cNSS↑	cAUC↑
ResNet-50	Sup.	0.517	0.439	−0.31	3.25	0.816
ViT-B/16	Sup.	0.568	0.490	1.30	3.83	0.871
DINO ViT-B/16	SSL	0.559	0.473	1.51	3.90	0.887
DINOv2 ViT-B/14	SSL	**0.586**	**0.514**	1.83	**4.19**	0.896
DINOv3 ViT-B/16	SSL	0.581	0.501	**2.11**	4.17	**0.901**

4.5. Backbone Ablation

Table 2 compares backbones using the FPN pixel decoder and flow matching enabled. All backbones are frozen.

Self-supervised backbones substantially outperform supervised ones. DINOv2 and DINOv3 achieve the strongest results overall, with DINOv2 leading on SS (0.586) and SemSS (0.514), while DINOv3 achieves the best conditional metrics cIG (2.11) and cAUC (0.901). We select DINOv3 as our default backbone for its stronger conditional saliency performance, which reflects per-step prediction quality. All ViT backbones outperform ResNet-50 by a wide margin (+0.042–0.069 SS), confirming that vision transformer features are better suited for scanpath prediction even when frozen.

4.6. Flow Matching Ablation

Table 3 isolates the effect of flow matching.

Table 3. **Effect of flow matching** (DINOv3 backbone, FPN pixel decoder). FM is discarded at inference.

Flow Matching	SS↑	ΔSS	SemSS↑	cIG↑	cNSS↑
✓	**0.581**	—	**0.501**	**2.11**	**4.17**
	0.518	−0.063	0.434	0.78	3.81

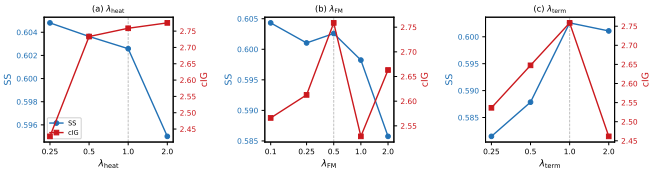


Figure 2. Sensitivity analysis of loss weights. (a) λ_{heat}: SS is stable while cIG increases with weight; 1.0 balances both. (b) λ_{FM}=0.5 provides optimal regularization. (c) λ_{term}=1.0 is critical for both SS and cIG.

Flow matching provides +0.063 SS improvement (0.518 → 0.581, +12.2% relative), despite being discarded at inference. The cIG improvement is particularly dramatic (0.78 → 2.11), indicating that FM substantially improves the informativeness of per-step fixation predictions by regularizing the decoder features to encode precise spatial coordinates.

4.7. Hyperparameter Sensitivity

We evaluate the sensitivity of FlowScan to key training hyperparameters: the loss balancing weights and the ground-truth heatmap scale σ.

Loss Weight Tuning. Figure 2 illustrates the impact of varying λ_{heat}, λ_{FM}, and λ_{term} on Sequence Score (SS) and Conditional Information Gain (cIG), sweeping each while holding the others at their default values (λ_{heat}=1.0, λ_{FM}=0.5, λ_{term}=1.0).

For the focal heatmap weight λ_{heat} (Fig. 2a), SS is stable across $[0.25, 1.0]$ (0.605–0.603) and drops slightly at 2.0 (0.595), while cIG increases monotonically (2.43 → 2.78), reflecting a trade-off between scanpath structure and per-fixation informativeness; λ_{heat}=1.0 provides the best balance.

For λ_{FM} (Fig. 2b), SS remains flat across $[0.1, 0.5]$ (≈ 0.604) and degrades at higher values (0.586 at 2.0), while cIG peaks sharply at 0.5 (2.76); this confirms that moderate flow matching regularization is beneficial, but overweighting it over-constrains the decoder.

The termination weight λ_{term} (Fig. 2c) shows the strongest sensitivity: SS drops from 0.603 at 1.0 to 0.582 at 0.25, and cIG degrades sharply at both extremes, indicating that accurate termination calibration is critical for both spatial and information-theoretic metrics.

Heatmap Sigma Tuning. The ground-truth heatmap is constructed by placing a Gaussian at the target fixation location. While HAT [30] uses a fixed 1° visual angle, we tune σ in pixel space. As shown in Table 4, $\sigma = 1.5$ yields the best results across all metrics. A smaller value ($\sigma = 1.0$) produces an overly peaked target with sparse gradient signal, while a larger value ($\sigma = 2.0$) blurs the target and reduces localization precision.

Table 4. **Effect of heatmap sigma** (DINOv3, deformable, FM).

σ	SS↑	SemSS↑	cIG↑	cNSS↑
1.0	0.595	0.528	2.518	4.758
1.5	**0.603**	**0.538**	**2.759**	**4.810**
2.0	0.595	0.523	2.535	4.710

4.8. Discussion

Why flow matching helps. The flow matching head provides an auxiliary gradient signal through the decoder features that is complementary to the heatmap focal loss. While the focal loss operates on a 2D spatial map and assigns most gradient to the peak location, the flow matching loss directly regresses fixation coordinates via a learned velocity field. This provides: (1) a smoother loss landscape around the target fixation, (2) direct backpropagation of coordinate error into the decoder representation, and (3) implicit enforcement of fixation locality (the velocity field must converge to a specific point). The result is a decoder representation that more precisely encodes "where to look next," improving heatmap quality even though the FM head is not used at inference.

Self-supervised features for visual search. DINOv3's features provide three advantages over supervised ResNet-50: (1) richer semantic correspondences that enable better task-target matching, as evidenced by higher SemSS; (2) emergent object boundary encoding [3] useful for foveal crop extraction; and (3) dense 768-dimensional patch tokens that encode contextual relationships across the scene.

Deformable pixel decoder. Comparing FlowScan with a deformable pixel decoder (SS = 0.603) to the FPN variant (SS = 0.581), the deformable decoder provides +0.022 SS (+3.8% relative). Multi-scale deformable attention enables cross-scale feature fusion that the lightweight FPN cannot achieve, leading to higher-quality P4 features for both foveal crops and dense heatmap prediction. The conditional saliency metrics show an even larger relative gain: cIG improves by +31% (2.11 → 2.76).

4.9. Qualitative Results

Figure 3 shows qualitative scanpath predictions on representative test images. For each example, we display the ground-truth scanpath (from one human subject) alongside predictions from FlowScan and HAT retrained. The target



Figure 3. **Qualitative scanpath predictions.** Each row shows a different test image with the search task indicated on the left. Columns show: one human ground-truth scanpath and model predictions. Green dashed boxes indicate the target object. Numbered circles show fixation order; arrows indicate saccade direction.

object is indicated by a green dashed bounding box, and fixations are numbered sequentially with connecting arrows showing the saccade trajectory.

5. Conclusion and Future Work

We presented FlowScan, a scanpath prediction model that builds on the HAT architecture with three modifications: a frozen self-supervised DINOv3 backbone, an adapted deformable pixel decoder with soft-argmax decoding, and a flow matching auxiliary head that regularizes the decoder representation. On COCO-Search18 target-present search, FlowScan achieves SS = 0.603 and SemSS = 0.538, outperforming HAT retrained on the same test split (SS = 0.575, SemSS = 0.499) across all five evaluation metrics.

Ablation studies confirm that each component contributes meaningfully. Self-supervised ViT backbones outperform supervised ResNet-50 by a wide margin (+0.064 SS), with DINOv3 providing the best conditional saliency metrics. Flow matching regularization yields +0.063 SS (+12.2% relative) despite being discarded at inference, demonstrating that auxiliary training objectives can substantially improve decoder representations without adding inference cost. The deformable pixel decoder provides an additional +0.022 SS (+3.8%) over the FPN variant through multi-scale feature fusion.

Several directions remain promising for future work: (1) extending to Target-Absent search, potentially with Inhibition-Of-Return mechanisms, (2) exploring stochastic generation via the flow matching head to produce diverse scanpath samples, (3) evaluating on additional eye-tracking datasets beyond COCO-Search18, (4) investigating whether the flow matching regularization principle transfers to other sequential prediction tasks, and (5) incorporating explicit, polar, space-dependent saccadic priors into FlowScan, building on recent evidence that both stimulus-center bias and saccade statistics are strongly space-dependent [1, 21, 22].

Acknowledgments

The authors acknowledge the use of generative AI tools for technical assistance in code implementation and linguistic refinement.

References

[1] Markus Bindemann. Scene and screen center bias early eye movements in scene viewing. *Vision Research*, 50(23):2577–2587, 2010. Vision Research Reviews. 8

[2] 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, 2013. 1, 2

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021. 1, 2, 7

[4] Yupei Chen, Zhibo Yang, Seoyoung Ahn, Dimitris Samaras, Minh Hoai, and Gregory Zelinsky. Coco-search18 fixation dataset for predicting goal-directed attention control. *Scientific Reports*, 11(1):8776, 2021. 1, 2, 5, 6

[5] Yupei Chen, Zhibo Yang, 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 Workshops (CVPRW)*, pages 5765–5774, 2022. 2, 6

[6] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002. 5

[7] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations (ICLR)*, 2024. 2, 3

[8] Richard Dewhurst, Marcus Nyström, Halszka Jarodzka, Tom Foulsham, Roger Johansson, and Kenneth Holmqvist. It depends on how you look at it: Scanpath comparison in multiple dimensions with multimatch, a vector-based approach. *Behavior research methods*, 44(4):1079–1100, 2012. 2

[9] Yini Fang, Jingling Yu, Haozheng Zhang, Ralf van der Lans, and Bertram Shi. Oat: Object-level attention transformer for gaze scanpath prediction. In *European Conference on Computer Vision*, pages 366–382. Springer, 2024. 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1, 3

[11] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022. 2

[12] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, 2001. 1

[13] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998. 1, 2

[14] Halszka Jarodzka, Kenneth Holmqvist, and Marcus Nyström. A vector-based, multidimensional scanpath similarity measure. In *Proceedings of the 2010 symposium on eye-tracking research & applications*, pages 211–218, 2010. 2

[15] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9404–9413, 2019. 5

[16] Matthias Kummerer, Thomas SA Wallis, Leon A Gatys, and Matthias Bethge. Understanding low-and high-level contributions to fixation prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 4789–4798, 2017. 1, 2, 5

[17] Peizhao Li, Junfeng He, Gang Li, Rachit Bhargava, Shaolei Shen, Nachiappan Valliappan, Youwei Liang, Hongxiang Gu, Venky Ramachandran, Yang Li, et al. Uniar: A unified model for predicting human attention and responses on visual content. *Advances in Neural Information Processing Systems*, 37:106346–106369, 2024. 1, 2

[18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 3

[19] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. Originally appeared as arXiv:2210.02747. 1, 2, 3

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. Originally appeared as arXiv:1711.05101. 6

[21] Rotem Mairon and Ohad Ben-Shahar. The polar saccadic flow model: Re-modeling the center bias from fixations to saccades. *Vision Research*, 228:108546, 2025. 8

[22] Rotem Mairon and Ohad Ben-Shahar. Stimulus center bias persists irrespective of its position on the display. *Journal of Eye Movement Research*, 18(6):77, 2025. 8

[23] 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*, pages 1441–1450, 2023. 1, 2, 6

[24] Takumi Nishiyasu and Yoichi Sato. Gaze scanpath transformer: Predicting visual search target by spatiotemporal semantic modeling of gaze scanpath. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 625–635, 2024. 2

[25] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 2, 3

[26] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 1, 2, 3

[27] Jeremy M Wolfe. Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review*, 1(2):202–238, 1994. 2

[28] Jeremy M Wolfe. Guided search 6.0: An updated model of visual search. *Psychonomic bulletin & review*, 28(4):1060–1092, 2021. 1, 2

[29] 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*, pages 193–202, 2020. 1, 2, 6

[30] 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*, pages 1683–1693, 2024. 1, 2, 3, 5, 6, 7

[31] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. Originally appeared as arXiv:2010.04159. 2, 3