TDYSN: A YOLO-Based Top Down Saliency Prediction Network

Can Mızraklı* and Tolga Kurtuluş Çapın†
Department of Computer Engineering, TED University, Ankara, Turkey
*can.mizrakli@tedu.edu.tr, †tolga.capin@tedu.edu.tr

***Abstract*—Visual saliency aims to predict the regions of an image most likely to attract human visual attention. While most saliency models assume free-viewing conditions, human attention is often shaped by explicit task goals. In this work, we address task-driven saliency prediction by proposing a model that conditions visual attention on natural-language task descriptions. The model produces task-dependent saliency maps that reflect how attention shifts under different viewing intents. Through quantitative and qualitative analysis, we show that incorporating explicit task semantics enables more faithful modeling of goal-directed visual attention.**

***Index Terms*—Visual Saliency, Top-Down Saliency, YOLO, Visual Attention, Gaze**

Code & data: https://github.com/canmizrakli/TDYSN

I. Introduction

Visual attention prediction is central to understanding how humans interpret visual scenes. Traditional bottom-up saliency methods focus on low-level visual information like color, edges, and contrast, assuming a common gaze pattern across observers. However, in real-world settings, attention is often task-driven: for example, searching for pedestrians while driving or locating specific objects in surveillance footage. These top-down goals fundamentally reshape how visual saliency should be modeled.

Task-conditioned saliency prediction is becoming increasingly important in domains such as autonomous driving, human–robot interaction, and assistive vision. However, existing approaches often fall short in two critical aspects: (1) they do not explicitly incorporate high-level task semantics into the saliency generation process, and (2) they lack a lightweight mechanism for combining such semantic cues with multiscale visual features. Approaches that rely solely on complex architectures (e.g., transformer-based solutions) tend to suffer from reduced spatial resolution and high computational cost, whereas others neglect task conditioning altogether and thus fail to capture goal-directed human attention patterns.

To address these limitations, we structure our study around the following research questions:

- **RQ1:** How can a pretrained object-detection backbone be leveraged to extract multiscale visual features that support task-oriented saliency modeling?
- **RQ2:** How can textual task definitions, encoded through sentence-level embeddings, be fused with spatial visual features to guide attention maps?
- **RQ3:** Can a model combining vision and language improve task-driven saliency prediction compared to traditional saliency models?

Figure 1 illustrates the motivation behind our work. Human gaze is not fixed but depends strongly on the task being performed; the same visual stimulus can yield different fixation patterns depending on the viewer's intent. The diagram highlights this task-dependency and shows how our model integrates the image stimulus with a natural-language task description to produce a task-conditioned saliency map. This provides a high-level view of how semantic intent shapes visual attention and motivates the architectural design introduced in Section III.

A. Contributions

Our work introduces **TDYSN**, a modular vision-language architecture for task-driven saliency prediction. Building upon the research questions outlined above, our main contributions are as follows:

- We introduce **TDYSN**, a modular top-down saliency prediction network that fuses YOLO-derived multiscale visual features with Sentence-BERT task embeddings via a transformer-based fusion block.
- We propose a *lightweight architectural design* for task-driven saliency, where semantic conditioning is integrated through a shallow fusion transformer used exclusively for cross-modal interaction, rather than through deep multimodal backbones.
- We validate our model through five standard saliency metrics (CC, KLDiv, SIM, NSS, AUC-Borji) and qualitative visualizations, and discuss its applicability in real-world, attention-aware vision systems.
- We provide an empirical analysis (including ablations and metric-category interpretation) showing that different architectural components impact distinct attention properties (spatial correlation vs. distributional alignment vs. saliency strength), clarifying what matters for task-driven saliency beyond a single-score view.

II. Related Work

Visual saliency aims to predict where humans fixate within a scene. Traditional bottom-up models rely on low-level cues such as color or contrast and assume task-free viewing, whereas top-down saliency reflects how goals and semantic intent shape attention. Task-driven saliency research therefore focuses on

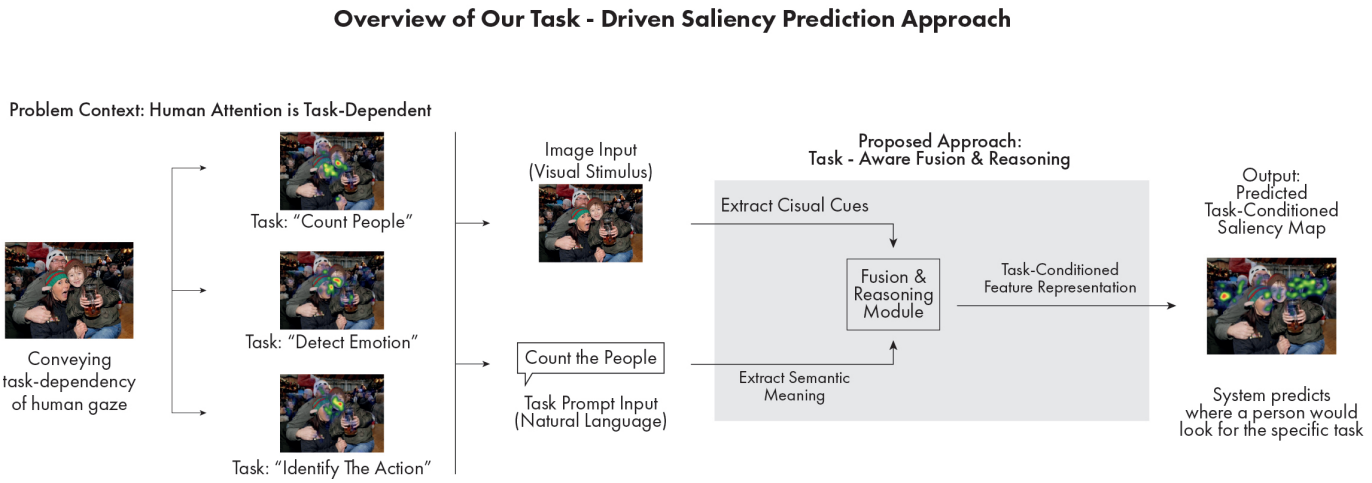

Figure 1. Conceptual overview of task-driven saliency prediction. Different task prompts lead to distinct human fixation patterns on the same image. We model this by fusing visual features extracted from the image with semantic representations of the task prompt, producing a task-based saliency map that reflects where a viewer would look given a specific goal.

how high-level intent influences gaze behavior, often using eye-tracking datasets whose fixation density maps (FDMs) provide probabilistic ground truth under different task conditions. This perspective motivates models that can integrate semantic cues with visual structure, forming the basis for the approaches reviewed below.

Research in saliency prediction has evolved from purely bottom-up methods to top-down approaches that incorporate task-specific and semantic cues. In this section, we categorize prior work into three main directions: (1) early bottom-up and task-free saliency models, (2) task-driven saliency prediction, and (3) vision–language fusion methods for attention modeling.

A. Bottom-Up vs. Task-Driven Saliency Models

Early saliency research primarily focused on bottom-up models that rely on low-level features such as color, texture, and contrast. These models assume a universal attention pattern, independent of viewer intent. Learning-based saliency models later showed that combining low-, mid-, and high-level cues learned from eye-tracking data improves fixation prediction over purely hand-crafted bottom-up cues [1]. However, such approaches still fail to capture task-based behaviors present in many real-world scenarios.

An early attempt at top-down saliency was made by Liang and Zhang [2], who introduced a KL-divergence–based formulation to identify regions contributing most to object recognition. Although their approach conditions attention on semantics indirectly through class distributions, it does not incorporate explicit task or goal definitions.

Top-down models address this limitation by conditioning attention on high-level intent or task definitions. For example, Murabito et al. [3] propose saliency maps that emerge from classification tasks, aligning attention with class-relevant regions. Similarly, Albayrak [4] explores task-conditioned attention using eye-tracking under various viewing goals. These models demonstrate that task semantics can significantly alter attention patterns.

Unlike these earlier approaches, our model explicitly incorporates semantic task information through a dedicated language encoder and conditions visual attention based on high-level intent, enabling more targeted and context-aware saliency predictions.

B. Saliency Prediction Using Transformers

Transformers have shown strong performance in saliency prediction by modeling global dependencies. Liu et al.'s Visual Saliency Transformer (VST) [5] uses patch-wise tokenization and multihead self-attention to capture long-range interactions, improving over convolutional baselines. While effective, these methods often require large-scale data and computation, making them less suitable for real-time applications.

Transformers have also been used in multimodal settings, where visual tokens are combined with textual or semantic tokens to guide spatial attention. This formulation enables cross-modal interactions, allowing linguistic cues to influence the weighting and interpretation of spatial features. Prior work in vision-language models has shown that appending or concatenating textual embeddings as additional tokens within a transformer sequence can effectively steer attention toward regions aligned with linguistic intent [6], [7].

Our work draws inspiration from this direction but simplifies the design by using a transformer encoder only for fusion, while relying on YOLO for efficient multiscale feature extraction. In contrast to large transformer-based saliency models that require extensive computation, our approach limits transformer usage to multimodal fusion, maintaining the contextual reasoning needed for task-driven attention without architectural complexity.

C. Vision–Language Fusion for Attention Modeling

Recent models have increasingly bridged vision and language to enable conditioned attention. Sentence-BERT [8] and other

pretrained language encoders convert textual prompts into dense semantic embeddings suitable for multimodal learning. Early work by Ramanishka et al. [9] demonstrated that natural language captions can effectively guide visual saliency, laying the foundation for vision–language grounded attention modeling.

Building upon this, MDETR [10] employs transformer-based architecture to align textual queries with visual regions, enabling end-to-end modulated object detection through vision-language alignment, though not directly optimized for saliency. More recently, TDiffSal [11] adopts a diffusion-based framework to generate saliency maps directly conditioned on textual descriptions, highlighting the growing relevance of text-driven saliency prediction.

Our approach builds upon these developments by hypothesizing that task-driven saliency can emerge naturally when semantic task intent and object-level visual features are jointly modeled within a unified transformer-based fusion space. Specifically, we posit that aligning high-level textual semantics with mid-level visual representations enables the network to emphasize task-relevant regions without the need for explicit supervision of attention boundaries. This hypothesis is inspired by cognitive findings suggesting that human attention is dynamically modulated by goal-related cues, our model mirrors this behavior computationally by linking linguistic task prompts to spatial feature activations extracted from an object detection backbone.

Unlike prior multimodal models such as MDETR and TDiffSal, which primarily target detection or rely on computationally heavy generative diffusion mechanisms, our method focuses on goal-specific saliency prediction by explicitly linking semantic task intent with visual attention. By fusing Sentence-BERT-derived task embeddings with YOLO-based visual features via a transformer, we establish an interpretable pathway between high-level task semantics and spatial attention patterns, enabling top-down, task-aware attention modeling without imposing assumptions tied to a specific detection or generative objective.

Taken together, these prior works collectively underscore the growing importance of integrating semantic task definitions with visual feature extraction for interpretable vision systems. Building on these observations, we introduce **TDYSN**, a top-down, language-conditioned saliency network designed to model task-driven visual attention. The proposed approach bridges the gap between human task understanding and machine attention mechanisms, offering both quantitative performance gains and qualitative interpretability in task-specific saliency prediction.

III. Proposed Method

We propose a top-down saliency prediction model that estimates spatial distributions of human visual attention under task-driven viewing conditions. By integrating a pre-trained YOLO-based architecture for feature extraction with a Sentence-BERT-based task encoder, our methodology bridges the gap between visual features and semantic task definitions. The YOLO backbone performs well at extracting visual cues, and the Sentence-BERT encoder transforms high-level text-based task descriptions into semantic embeddings. Then the transformer-based fusion module processes these two components together to create a connection between extracted features based on tasks. Consequently, the model benefits from the strengths of both object detection and contextual understanding, resulting in more accurate task-specific saliency maps.

A. Design Hypotheses

The architectural design of TDYSN is guided by the following hypotheses, which aim to answer the research questions stated in the Introduction:

- **H1:** Incorporating an object detection backbone such as YOLO enables the model to extract semantically rich and spatially precise visual features that align with human gaze patterns under specific tasks.
- **H2:** Encoding high-level task definitions through Sentence-BERT embeddings provides meaningful semantic cues that can steer saliency prediction toward goal-relevant regions.
- **H3:** A fusion of visual and textual representations will lead to improved task-driven attention alignment, enhancing both performance metrics (e.g., NSS, AUC-Borji) and interpretability of the resulting saliency maps.

B. Model Architecture

The architecture consists of a YOLO-based backbone for visual feature extraction, an FPM (1×1 conv projection), a task encoder to generate semantic embeddings, a transformer fusion module that combines visual and semantic features, and a final saliency decoder that produces the output saliency maps (see Figure 2).

1) YOLO Backbone: The backbone employs a pre-trained YOLO model [12], [13] to extract multiscale visual features from input images. YOLO's hierarchical architecture captures both low-level cues such as edges and textures, and high-level semantic structures like objects and contextual relationships, factors that closely align with human visual attention mechanisms. By leveraging its pretrained detection layers, our model benefits from a strong object-centered prior that enhances scene understanding while reducing the need for extensive training from scratch. We specifically utilize the earlier layers of YOLO, optimized for feature extraction rather than classification, to provide efficient and semantically rich representations for task-driven saliency prediction.

2) Feature Projection Module (FPM): The Feature Projection Module applies a single 1×1 convolution to the high-dimensional feature maps produced by the YOLO backbone, reducing their channel dimensionality from 512 to 128 while preserving spatial resolution. This operation serves to distill the most informative visual cues into a more compact representation that retains critical spatial details. The reduced dimensionality minimizes computational cost and memory footprint during transformer fusion, ensuring that only the most salient and task-relevant features are carried forward. By simplifying the feature space without losing semantic richness,

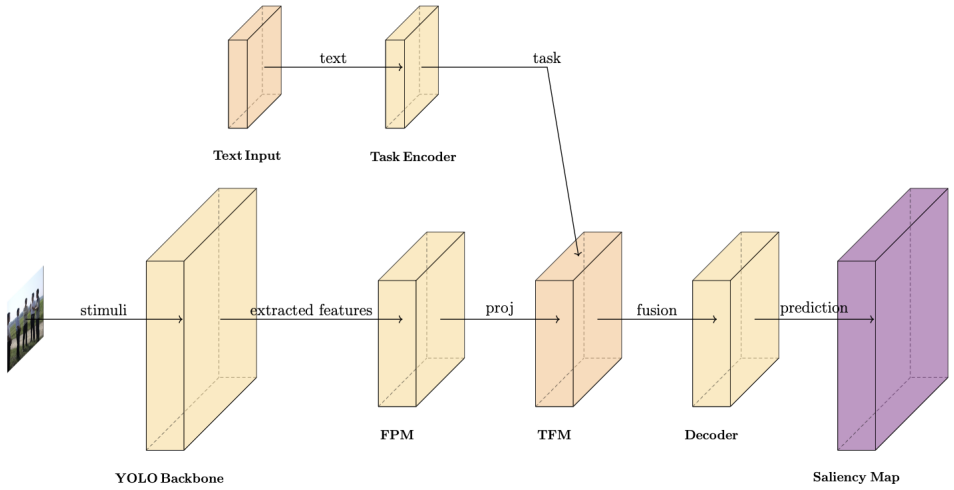

Figure 2. Overview of the model architecture.

the FPM enables efficient cross-modal fusion and contributes to the overall effectiveness of the saliency decoder.

3) Task Encoder: To incorporate task-specific context, the task encoder converts textual task descriptions into dense semantic embeddings using Sentence-BERT [8]. This model was chosen for its ability to capture sentence-level meaning efficiently, making it well suited for short, natural-language prompts that define viewing goals. The resulting embeddings are then refined through a linear transformation, producing a compact representation that aligns dimensionally with the visual feature space. This alignment bridges the language–vision gap, allowing semantic task intent to guide spatial attention toward regions in the image that are most relevant to the described goal.

4) Transformer Fusion Module (TFM): The Transformer Fusion Module unifies the visual and semantic representations to enable task-aware attention. Feature maps from the FPM are first flattened into a sequence of visual tokens, while the encoded task embedding expressing the high-level intent of the task is appended as an additional token. This combined sequence is processed by a transformer encoder that leverages self-attention to model the interactions between task semantics and spatial image features. Through this mechanism, the module learns to highlight regions that are most relevant to the specified task, effectively translating semantic guidance into spatial focus. The fused representation is then reshaped back into spatial form and forwarded to the saliency decoder for reconstruction.

5) Saliency Decoder: The saliency decoder reconstructs a dense attention map from the fused multimodal representation produced by the transformer. It consists of a shallow stack of convolutional and upsampling layers that progressively restore spatial resolution while refining feature coherence. This design preserves fine spatial detail without introducing unnecessary depth or computational cost. A final sigmoid activation normalizes the output to the $[0, 1]$ range, converting the response into a probabilistic saliency distribution. The resulting single-channel map highlights regions most likely to attract human attention under the given task condition, providing a directly interpretable and task-driven output.

Table I
IMPLEMENTATION DETAILS

Component	**Key Settings**
Backbone	YOLOv5-su (Ultralytics) first 10 layers (SPPF output, 512-ch)
Feature Projection Module (FPM)	1×1 conv: 512→128 channels output $[B, 128, H, W]$
Task Encoder	SentenceTransformer (all-MiniLM-L6-v2): 384→64-dim
Transformer Fusion	1-layer Transformer Encoder $d_{\text{model}} = 128,\ n_{\text{head}} = 4$ + 1 task token appended
Saliency Decoder	Conv $3 \times 3 \rightarrow 64 \rightarrow$ Deconv×2→32 $\rightarrow$ Deconv×2→1 Sigmoid $\rightarrow [B, 1, H, W]$
Reproducibility	Random seed 42 for dataset splitting

C. Dataset & Preprocessing

The dataset used here was originally collected by Albayrak [4]. Large-scale saliency benchmarks such as SALICON have enabled data-driven saliency modeling by providing scalable human attention annotations (via mouse-tracking proxies) [14]. Our dataset, in contrast, provides eye-tracking fixation density maps under explicit task conditions, where each task is defined by a natural-language viewing goal. It comprises original stimuli images and, for each of four task conditions, two subfolders: `fdm` (grayscale eye-tracker saliency maps) and `heatmap` (RGB overlays on the stimuli).

In our experiments, we use only the raw grayscale maps as ground truth, yielding a total of 1,968 image–saliency pairs across the four tasks. We split these into:

- 70% train (1,377 pairs)
- 15% validation (295 pairs)
- 15% test (296 pairs)

All splits are generated with a fixed random seed (42) to ensure reproducibility.

Preprocessing and augmentations are applied identically to each image–map pair:

- **Resize:** Stimuli to 384×384 px. Ground-truth FDMs are recorded at 48×48; we keep them at that size and later down-sample predictions to 48×48 before computing the loss.
- **Normalization:** scale image pixel values to $[0, 1]$ and convert to tensors.
- **Paired augmentations:** random horizontal flip (probability 0.5) and random rotation within $\pm 10°$, applied the same way to the image and its corresponding map to preserve spatial correspondence.

Ground-truth fixation density maps are recorded at a reduced spatial resolution. We keep them at this resolution and down-sample predictions accordingly before computing the loss, consistent with standard saliency evaluation, where fixation maps are defined as smoothed density maps [15].

D. Training Configuration

We train our TDYSN model for 50 epochs. The optimizer is Adam with a constant learning rate of 1×10^{-4} and no additional weight decay. We use a batch size of 8 and reset gradients at each step to stabilize multimodal fusion on a modest dataset without overfitting.

The loss is a combined saliency objective:

$$\mathcal{L}(S, \hat{S}) = \alpha \, \mathrm{KL}(S \,\|\, \hat{S}) \; + \; \beta\big(1 - \mathrm{CC}(S, \hat{S})\big),$$

with weighting factors $\alpha = \beta = 1.0$. Here, KL stands for the *Kullback–Leibler divergence* [2] [15], which measures the discrepancy between the predicted saliency distribution $\hat{S}$ and the ground-truth distribution S, and CC stands for the *Pearson correlation coefficient* [15], interpreted as the cosine of the angle between the predicted and ground-truth saliency maps.

At each forward pass, the model outputs a 96×96 saliency map, which we bilinearly downsample to 48×48 to match the ground-truth eye-tracker maps before computing the loss. We log both batch-level and epoch-average loss values to monitor convergence (Figure 3).

IV. Results

A. Quantitative Evaluation

To evaluate the performance of our TDYSN model, we computed five widely adopted saliency metrics (Figure 4 & Table II) across the validation set: Pearson's Correlation Coefficient (CC) [15], Kullback-Leibler Divergence (KLDiv) [2], [15], Similarity (SIM) [15], Normalized Scanpath Saliency (NSS) [15], and AUC-Borji [16]. These metrics reflect different aspects of saliency quality: CC measures spatial correlation, KLDiv measures distribution divergence, SIM captures global map similarity, NSS evaluates alignment with fixation points, and AUC-Borji measures discriminative power between salient and non-salient regions (Figure 4).

Table II
Validation vs. Test performance across five saliency metrics.

Metric	Validation	Test
CC	0.6516	0.6423
KLDiv	0.9132	0.9270
SIM	0.5006	0.5010
NSS	3.6885	3.4583
AUC-Borji	0.9549	0.9486

As shown in Table II, validation and test metrics are closely aligned across all measures, with differences typically below 6%, indicating strong generalization and no signs of overfitting.

Our model achieves a NSS of 3.53 and AUC-Borji of 0.95, both of which are highly competitive results in the literature. This comparison to MIT/Tübingen Saliency Benchmark [17] is later discussed in the Comparison to Literature section (VI-B).

We follow the saliency evaluation methodology outlined by Kümmerer et al. [15], which emphasizes separating models, maps, and metric computation. Although our dataset differs from MIT300 and SALICON, the results demonstrate strong generalization and alignment with benchmark expectations.

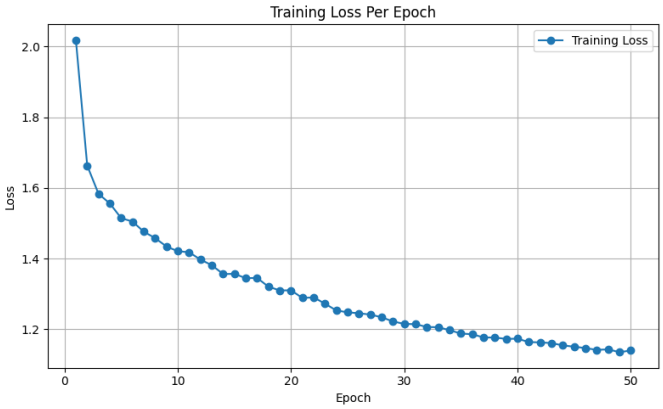

Figure 3. Training loss per epoch for 50 epochs.

B. Training Convergence

Figure 3 illustrates the training loss across 50 epochs. The model shows a consistently decreasing trend in loss, dropping from an initial value above 1.65 to approximately 1.18 by the final epoch. This steady decline suggests effective learning throughout training. Importantly, the curve exhibits no significant oscillations or spikes, indicating there is no notable instability or overfitting. The rate of improvement slows down around epoch 30, suggesting convergence as the model begins to saturate in performance.

Overall, the smooth downward trajectory of the loss curve confirms that the training process is both stable and efficient, validating our architectural and optimization choices.

C. Metric Trends Over Time

Figure 4 shows the trends of five validation metrics: NSS, AUC-Borji, CC, SIM, and KL divergence, over the 50 training epochs.

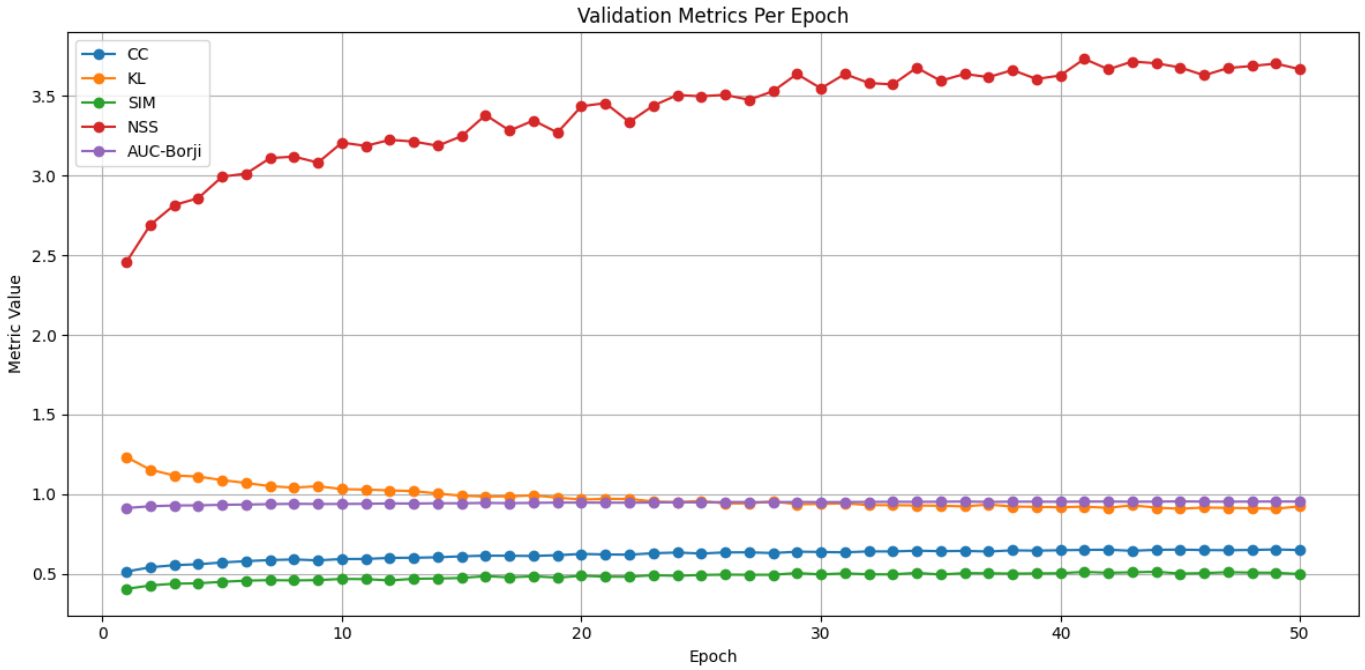

Figure 4. Validation metrics (CC, KL, SIM, NSS, AUC–Borji) over 50 epochs.

The trend that stands out the most is the steady and consistent increase in NSS, which starts around 2.65 and rises to 3.75 by epoch 50. This indicates that the model becomes increasingly aligned with human fixation patterns as training progresses. Similarly, AUC-Borji shows a gradual upward trend, improving from approximately 0.91 to 0.95, suggesting enhanced capability to rank salient regions correctly.

The CC and SIM metrics, also show incremental increase, indicating better global structure matching. Meanwhile, KL divergence steadily decreases from about 1.15 to below 0.95, reflecting reduced distributional mismatch between predicted and ground-truth saliency maps.

Overall, the trends across all metrics confirm that the model not only converges, but also generalizes better with longer training, with NSS and AUC-Borji showing the most significant improvements.

D. Qualitative Visualization

To better understand the task-driven behavior of our TDYSN model, we present saliency predictions on four different sample images under distinct task conditions. Each row in Figure 5 corresponds to a different task prompt, showing the ground-truth saliency map (FDM), and the predicted saliency map.

As seen in the results, the predicted saliency maps shift focus based on the given task prompt, demonstrating the model's ability to incorporate task semantics into spatial attention. The inclusion of the original stimuli images with ground truths as heatmaps on them helps illustrate how these attention shifts are contextually grounded in the visual scene, further validating the task-driven behavior of our model.

V. Explainability and Interpretability

Beyond predictive performance, task-driven saliency models must provide transparent and interpretable reasoning to be useful in human-centered vision systems [18], [19]. In this work, explainability is addressed through the way task information is incorporated into the model, rather than as a separate analysis step. TDYSN enables interpretability through explicit semantic conditioning, attention-based multimodal fusion, and human-aligned output representations, consistent with recent trends in interpretable vision–language modeling [20].

A. Task-Conditioned Causal Interpretability

A key explainability property of TDYSN is its explicit conditioning on natural-language task descriptions. Given the same visual stimulus, varying the task prompt leads to consistent and systematic variations in the predicted saliency distributions, as demonstrated qualitatively in Figure 5. Rather than producing radically different attention maps, the model exhibits task-dependent modulation of spatial focus, reflecting shifts in emphasis toward regions that are relevant to the specified goal. Similar task-dependent modulation effects have been observed in vision–language models when semantic inputs are varied under controlled conditions [20].

This behavior supports a counterfactual interpretation of the model's predictions: changing the task description while keeping the image fixed results in structured and predictable changes in spatial attention. Such task-conditioned saliency variations indicate that the model's predictions are influenced by semantic task information beyond low-level visual cues, while remaining grounded in the same underlying visual scene. This perspective is consistent with prior work on causal and

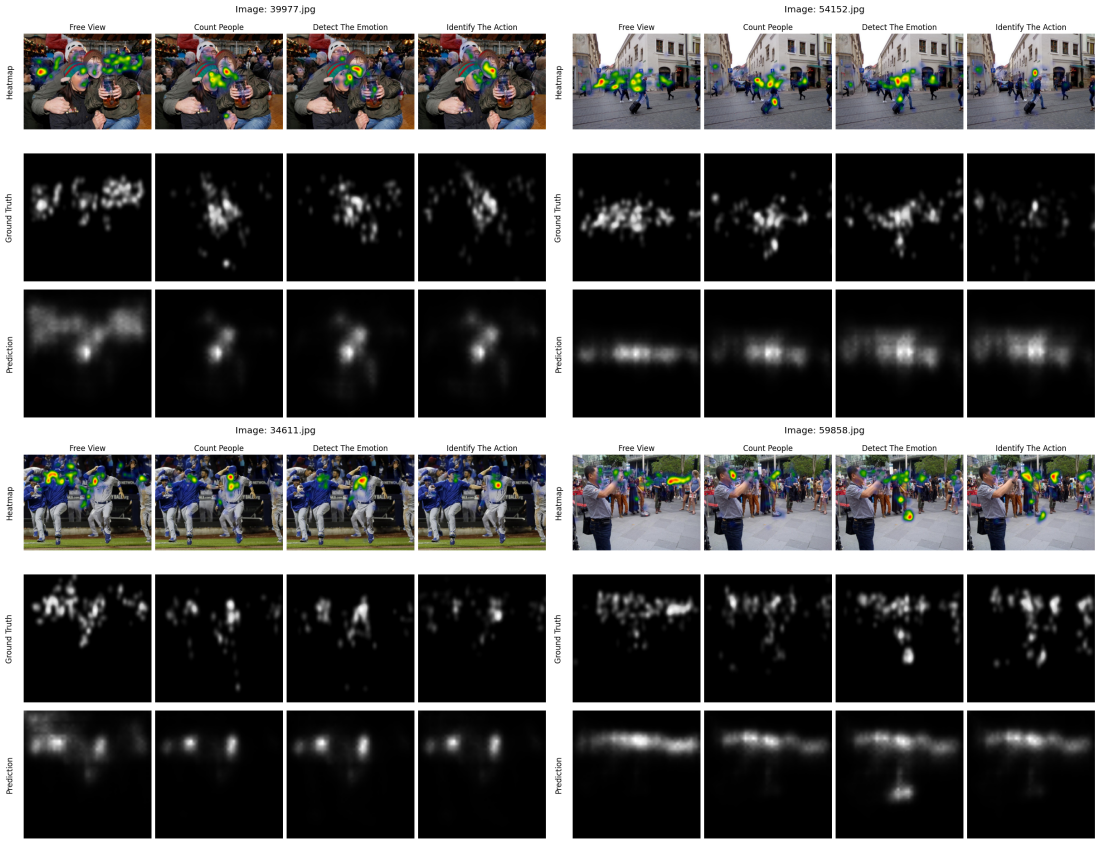

Figure 5. Example saliency predictions for four images under different task conditions. For each image, the top row shows the input image with task-specific attention overlays, the middle row shows the ground-truth fixation density maps, and the bottom row shows the predicted saliency maps.

counterfactual analysis in vision-language transformers, where semantic perturbations are used to probe attention and internal reasoning pathways [20].

B. Intrinsic Explainability via Transformer Fusion

Interpretability in TDYSN is further supported by its transformer-based fusion mechanism. The encoded task embedding is appended as a dedicated *task token* within the transformer input sequence, alongside spatial visual tokens derived from the YOLO backbone. Through self-attention, this task token can modulate the weighting of spatial features, enabling semantic intent to influence which regions receive higher emphasis in the fused representation. In this sense, the fusion transformer provides an *architecturally grounded* interface between language and spatial features, consistent with approaches that promote interpretability through structured attention mechanisms [21].

At the same time, we do not assume that raw attention weights are, by themselves, a complete or universally faithful explanation of model decisions. Instead, we treat attention as an *intrinsic diagnostic signal* that reveals how task information is propagated through the fusion module. Concretely, we visualize and analyze (i) the attention distribution from the task token to spatial tokens (averaged across heads), and (ii) how this distribution changes under counterfactual task prompts while keeping the input image fixed. If task-token attention shifts systematically toward task-relevant regions and these shifts covary with the corresponding changes in the predicted saliency map, this supports the interpretation that the fusion module is using task semantics in a structured and traceable way. This usage aligns with the broader view that attention can be informative for interpretability when it is semantically grounded and evaluated under controlled interventions, rather than presented as a standalone explanation [20].

C. Saliency Maps as Human-Aligned Explanations

The output of TDYSN consists of dense saliency maps aligned with human fixation density maps collected via eye tracking. These maps represent probabilistic predictions of where humans are likely to attend under a given task and therefore serve as naturally interpretable explanations of the model's behavior.

Quantitative improvements in NSS and AUC-Borji further support the claim that the predicted saliency distributions align well with human attention patterns, even when task-dependent differences are subtle. Such metrics have been widely adopted as objective measures of human-aligned explainability in saliency and attention modeling [22].

By framing saliency prediction itself as an explanation task, TDYSN aligns model interpretability with human perceptual

behavior, bridging the gap between algorithmic attention mechanisms and human-understandable visual reasoning. This perspective is consistent with recent human-centered XAI frameworks that emphasize alignment with human cognitive processes as a core requirement for trustworthy visual AI systems [18], [19].

D. Summary

Overall, TDYSN incorporates explainability at multiple levels: task conditioning provides causal interpretability through controlled semantic modulation, transformer fusion enables intrinsic attention-based reasoning, and saliency outputs offer human-aligned explanations. This combination positions TDYSN as an interpretable vision–language model, while acknowledging the inherent subtlety of task-driven attention shifts in complex visual scenes, as highlighted in recent analyses of multimodal explainability [18].

VI. Discussion

A. Interpretation of Results

The results presented in Section 4 confirm that our TDYSN model is capable of learning highly accurate, task-dependent saliency representations. The model achieved a NSS score of 3.53 and an AUC-Borji score of 0.9489, both of which indicate strong alignment with human fixation patterns. These metrics increased steadily across the 50 training epochs, suggesting stable convergence and continuous learning of task-relevant visual features. The upward trajectory in NSS is particularly notable, as it directly measures how well predicted saliency values align with human fixations at discrete locations. The consistent improvement in AUC-Borji further reinforces the model's ability to discriminate between salient and non-salient regions.

The CC and SIM scores, which are 0.6433 and 0.5112 respectively, reflect improved spatial structure preservation and histogram overlap, which are important for ensuring that predicted saliency maps capture both local focus and global attention cues. Meanwhile, KL divergence dropped steadily to 0.9239, reflecting reduced distributional mismatch between predictions and ground truth over time. Together, these trends indicate that the model not only converges smoothly but also generalizes its attention mapping across tasks and image types.

B. Comparison to Literature

We emphasize that TDYSN is trained and evaluated on a *task-conditioned* eye-tracking dataset, whereas widely used saliency benchmarks such as MIT300 report results primarily under *free-viewing* conditions and employ different stimuli, annotation protocols, and dataset biases. In addition, these benchmarks are designed to evaluate bottom-up, task-agnostic saliency, while TDYSN explicitly models *top-down, task-driven* attention.

As a result, the numerical values reported below should not be interpreted as a direct or competitive benchmark comparison. Instead, they are provided as a contextual reference to situate the scale of TDYSN's performance within the broader saliency literature.

Table III
Performance comparison between TDYSN, EML-Net, and the Gold Standard from MIT/Tübingen Saliency Benchmark [17].

Model	NSS	AUC	CC	KLDiv	SIM
TDYSN (Ours)	3.53	0.9489	0.6433	0.9239	0.5112
EML-Net	2.05	0.8660	0.8860	0.5200	0.7800
Gold Standard	3.14	0.9341	0.9828	0.0602	0.8992

According to the MIT/Tübingen Saliency Benchmark [17], the Gold Standard achieves an NSS of 3.14 and an AUC-Borji of 0.9341 on the MIT300 dataset, serving as a reference point for high-performing general-purpose saliency models under free-viewing conditions. We also include EML-Net [23], a strong CNN-based model evaluated on free-viewing benchmarks, which reports an NSS of 2.05 and an AUC of 0.866.

Within this contextual framing, TDYSN achieves NSS and AUC-Borji values that fall within the range typically associated with strong human-fixation alignment, despite being trained under task-conditioned viewing. At the same time, differences in CC and SIM reflect known trade-offs between task-driven and free-viewing saliency modeling. Metrics such as CC and SIM favor smoother and more globally correlated attention distributions, whereas task-conditioned saliency often yields sharper, semantically focused maps that improve fixation-based metrics such as NSS and AUC-Borji. This behavior is consistent with prior observations that goal-directed attention redistributes saliency mass toward task-relevant regions rather than globally salient image structures.

Beyond general-purpose saliency models, it is also informative to relate TDYSN to recent vision–language architectures such as MDETR and TDiffSal. MDETR [10] aligns textual queries with object-level visual regions for modulated detection, but it operates at the bounding-box level and is not designed to generate dense fixation distributions. TDiffSal [11], in contrast, produces text-conditioned saliency maps using a diffusion-based generative process, which entails substantially higher computational complexity. Within this landscape, TDYSN can be viewed as a discriminative alternative: by fusing Sentence-BERT task embeddings with YOLO-derived multiscale features through a transformer, it yields task-conditioned saliency maps that align well with human fixations without relying on large-scale generative modeling. We do not include numerical comparisons for MDETR [10] or TDiffSal [11] in Table III, as these methods are evaluated under different task formulations, output representations, and evaluation protocols, making direct metric-level comparison impossible.

C. Ablation Study

The ablation study is conducted to analyze how individual architectural components contribute to different properties of task-driven visual attention, rather than to optimize a single evaluation metric. Following prior saliency evaluation protocols, we group the metrics into three functional categories: (i) *spatial*

Table IV
ABLATION STUDY ON TDYSN. ARROWS INDICATE WHETHER HIGHER (↑) OR LOWER (↓) VALUES ARE BETTER.

Model	CC ↑	KL ↓	SIM ↑	NSS ↑	AUC-Borji ↓
Full TDYSN	0.6806	0.8641	0.5388	3.8449	-0.9264
w/o Task	0.6533	0.9563	0.5275	3.6965	-0.9005
w/o Transformer	0.6894	0.8700	0.5556	3.9790	-0.9197
w/o SBERT	0.6809	0.9218	0.5547	3.9762	-0.9079
w/o FPM	0.6886	0.8621	0.5557	4.0152	-0.9098

correlation with human attention (CC), (ii) *distributional alignment* (KL divergence and AUC-Borji), and (iii) *saliency strength* (NSS and SIM).

Removing task conditioning results in a consistent degradation across all metric categories, confirming that explicit task input is the primary driver of task-specific attention modulation in TDYSN. This validates the core premise of the proposed framework.

In contrast, removing the transformer fusion module, SBERT-based semantic encoding, or the FPM leads to clear metric-dependent trade-offs. While saliency strength metrics (NSS, SIM) may increase in these ablated variants, distribution-level alignment and robustness metrics (KL divergence and AUC-Borji) exhibit consistent degradation or instability relative to the full model, indicating reduced global coherence and semantic faithfulness of the predicted attention distributions.

These results suggest that the transformer, SBERT, and multi-scale feature fusion components act as regularizing mechanisms that promote coherent, semantically grounded, and globally consistent attention under varying task prompts, rather than maximizing sharp saliency responses in isolation.

D. Limitations

Despite the strong quantitative and qualitative results, several limitations remain. First, the dataset used in this study comprises 1,968 image–task pairs across four fixed task categories. While sufficient to demonstrate the feasibility and effectiveness of task-driven saliency modeling, this limited scale and semantic diversity may constrain the model's ability to generalize to unseen task formulations, open-ended natural language prompts, or domain-shifted visual content.

Second, although we report an ablation study that analyzes the contribution of key architectural components, the analysis remains metric-driven and task-aggregated. We do not explicitly examine task-specific ablation effects (e.g., how different components influence performance under individual task types), nor do we analyze cross-task transfer behavior. As a result, finer-grained insights into how architectural elements support specific task semantics remain unexplored.

Third, cross-dataset generalization could not be evaluated on benchmarks such as SALICON or MIT300 due to the lack of task-conditioned eye-tracking annotations on those datasets. Consequently, while our results demonstrate strong alignment with human attention under controlled task settings, they do not directly establish robustness across heterogeneous saliency benchmarks or free-viewing scenarios. As saliency evaluation is known to be sensitive to dataset bias and metric choice, absolute performance comparisons across datasets should therefore be interpreted with caution [16].

E. Future Work

Future work will focus on extending both the empirical scope and analytical depth of TDYSN. A primary direction is the expansion of the existing dataset through additional eye-tracking data collection, with the goal of increasing task diversity and improving generalization to more complex or compositional task descriptions.

From a modeling perspective, a future work is to extend the current ablation analysis by conducting task-specific and cross-task studies, enabling a deeper understanding of how individual components contribute under different semantic conditions. This includes examining whether components such as the transformer fusion module or the task encoder provide consistent benefits across all task types or primarily support certain forms of task-driven attention.

Architecturally, future work will explore multi-layer transformer fusion, task-adaptive decoding strategies, and alternative language encoders to better capture nuanced task semantics. Finally, we aim to investigate lightweight cross-dataset evaluation strategies, such as weakly supervised transfer or synthetic task conditioning, to assess robustness beyond the current dataset without requiring full task-specific eye-tracking annotations.

Together, these directions aim to improve the scalability, generalization, and interpretability of task-driven saliency models, enabling more robust deployment in real-world, human-centered vision systems.

VII. CONCLUSION

We presented TDYSN, a top-down saliency prediction network that couples a pre-trained YOLO backbone with a 1×1 Feature Projection Module, a Sentence-BERT task encoder, and a transformer-based fusion block. On a task-oriented eye-tracking dataset, TDYSN attains an NSS of 3.53 and AUC-Borji of 0.9489. While results from MIT/Tübingen Benchmark and prior free-viewing models are not directly comparable due to different datasets and protocols, these values provide a contextual indication that TDYSN achieves strong human-fixation alignment despite task conditioning and a lightweight fusion design. Consistently rising metric curves and qualitative visualizations confirm that the model not only converges with stability but also shifts attention in line with the provided task prompts, which proves that integrating high-level semantics with multi-scale visual cues is crucial for goal-directed saliency.

Despite these encouraging results, TDYSN is currently bounded by the scale and semantic breadth of the available data and has yet to be validated across heterogeneous saliency benchmarks. Future work will therefore focus on expanding and diversifying the dataset, performing cross-dataset studies, and modifying the architecture for enhanced performance. We believe these directions will extend TDYSN's applicability to real-world settings ranging from assistive vision and human–robot

interaction to attention-aware image understanding, and will stimulate further research at the intersection of visual attention, language grounding, and lightweight detection backbones.

REFERENCES

[1] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *2009 IEEE 12th international conference on computer vision*. IEEE, 2009, pp. 2106–2113.

[2] J. Liang and Y. Zhang, "Top down saliency detection via kullback-leibler divergence for object recognition," in *2015 International Symposium on Bioelectronics and Bioinformatics (ISBB)*, 2015, pp. 200–203.

[3] F. Murabito, C. Spampinato, S. Palazzo, D. Giordano, K. Pogorelov, and M. Riegler, "Top-down saliency detection driven by visual classification," *Computer Vision and Image Understanding*, vol. 172, pp. 67–76, 2018.

[4] D. Albayrak, "A study of visual saliency for free-viewing and task-oriented condition," Master's thesis, TED University, 2020.

[5] N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han, "Visual saliency transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 4722–4732.

[6] Y. Wen, Q. Cao, Q. Fu, S. Mehta, and M. Najibi, "Efficient vision-language models by summarizing visual tokens into compact registers," *arXiv preprint arXiv:2410.14072*, 2024.

[7] C. Neo, L. Ong, P. Torr, M. Geva, D. Krueger, and F. Barez, "Towards interpreting visual information processing in vision-language models," *arXiv preprint arXiv:2410.07149*, 2024.

[8] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[9] V. Ramanishka, A. Das, J. Zhang, and K. Saenko, "Top-down visual saliency guided by captions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[10] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion, "Mdetr-modulated detection for end-to-end multi-modal understanding," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1780–1790.

[11] N. Zhang, M. Xiong, D. Zhu, K. Zhu, and G. Zhai, "Tdiffsal: Text-guided diffusion saliency prediction model for images," in *International Conference on Pattern Recognition*. Springer, 2024, pp. 15–31.

[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[13] G. Jocher and Ultralytics, "Ultralytics YOLOv5," https://github.com/ultralytics/yolov5, 2020, version 7.0.

[14] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "Salicon: Saliency in context," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1072–1080.

[15] M. Kummerer, T. S. Wallis, and M. Bethge, "Saliency benchmarking made easy: Separating models, maps and metrics," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 770–787.

[16] A. Borji, D. N. Sihite, and L. Itti, "Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 55–69, 2012.

[17] M. Kümmerer, T. Wallis, and M. Bethge, "Mit/tübingen saliency benchmark," 2024, https://saliency.tuebingen.ai/results.html.

[18] M. Vatsa, A. Jain, and R. Singh, "Adventures of trustworthy vision-language models: A survey," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 20, 2024, pp. 22 650–22 658.

[19] U. Ehsan and M. O. Riedl, "Human-centered explainable ai: Towards a reflective sociotechnical approach," in *International Conference on Human-Computer Interaction*. Springer, 2020, pp. 449–466.

[20] X. Yang, H. Zhang, G. Qi, and J. Cai, "Causal attention for vision-language tasks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9847–9857.

[21] S. Chen, M. Jiang, J. Yang, and Q. Zhao, "Air: Attention with reasoning capability," in *European Conference on Computer Vision*. Springer, 2020, pp. 91–107.

[22] G. Liu, J. Zhang, A. B. Chan, and J. Hsiao, "Human attention-guided explainable ai for object detection," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 45, no. 45, 2023.

[23] S. Jia and N. D. Bruce, "Eml-net: An expandable multi-layer network for saliency prediction," *Image and vision computing*, vol. 95, p. 103887, 2020.