UniAR: Unifying Human Attention and Response Prediction on Visual Content

Peizhao Li*†[1], Junfeng He*‡[2], Gang Li*‡[2], Rachit Bhargava[2], Shaolei Shen[2], Nachiappan Valliappan[2], Youwei Liang†[3], Hongxiang Gu[2], Venky Ramachandran[2], Golnaz Farhadi[2], Yang Li[2], Kai J Kohlhoff[2], and Vidhya Navalpakkam[2]

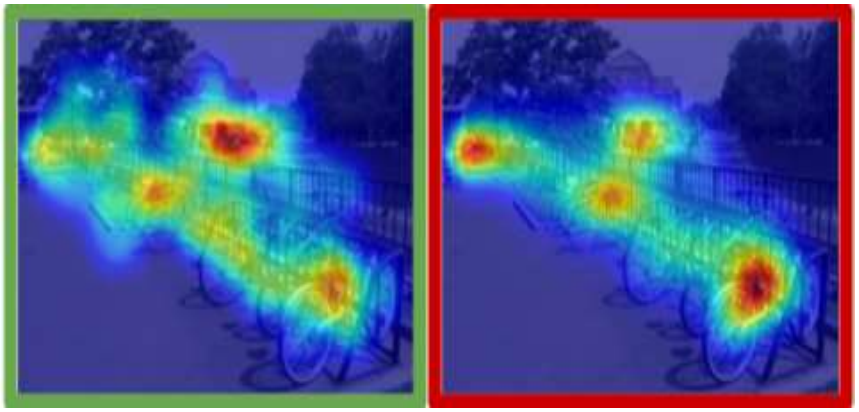
[1]Brandeis University
[2]Google Research
[3]University of California San Diego

2024.4.8

赖彦涛

1. Introduction

1 saliency heatmap



2 importance heatmap

3 scanpath prediction

4 Subjective rating prediction

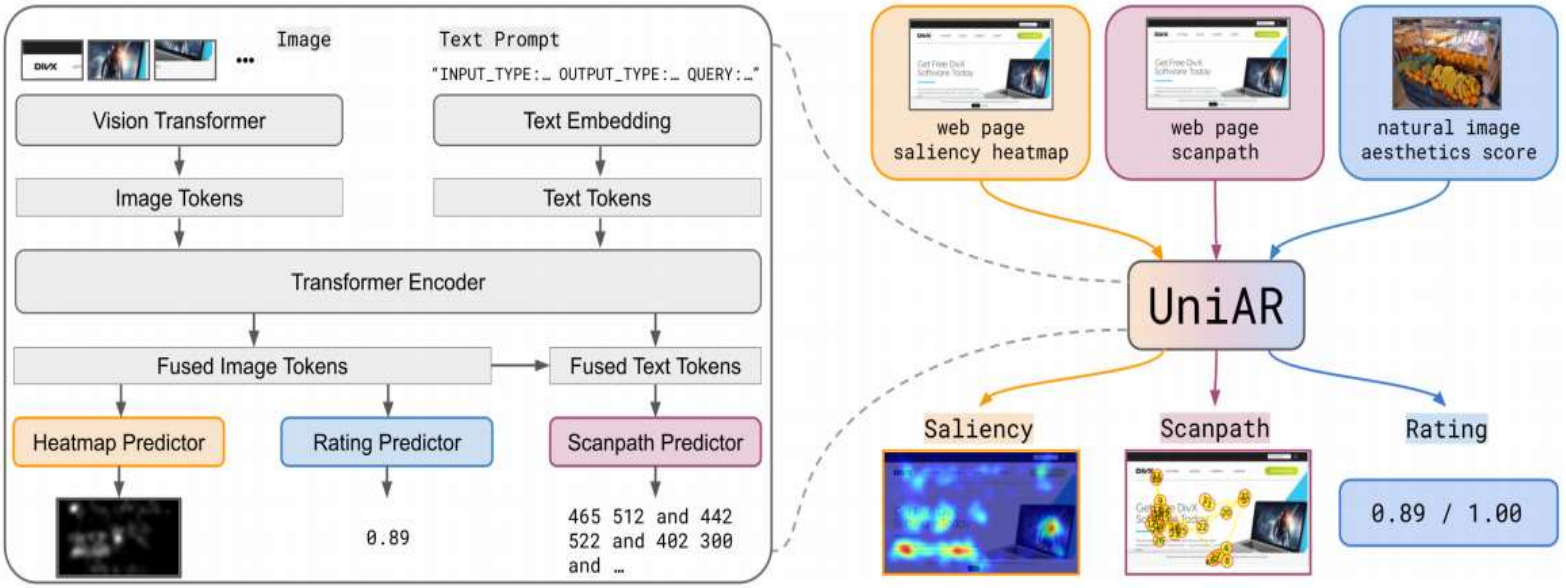
1. Introduction

Motivation

Progress in human behavior modeling involves under standing both implicit, early-stage perceptual behavior such as human attention and explicit, later-stage behavior such as subjective ratings/preferences. Yet, most prior research has focused on modeling implicit and explicit human behavior in isolation.

This paper proposed UniAR, a unified multimodal transformer model to predict both implicit and explicit human behavior across diverse types of visual content and tasks

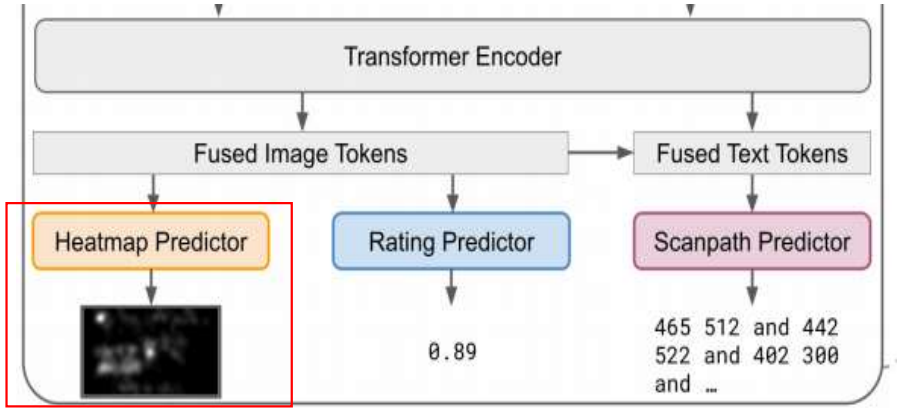
2. Methodology



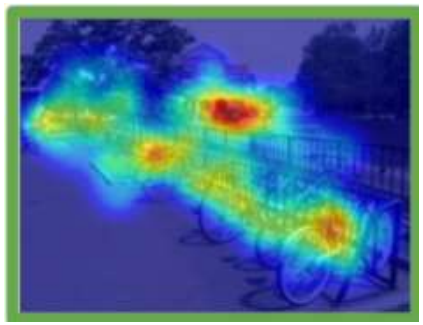
This paper adopt a multimodal encoder-decoder transformer model to unify the various human behavior modeling tasks. The model takes two types of inputs: an image and a text prompt. Its architecture comprises several components: a Vision Transformer model for image encoding, a word embedding layer to embed text tokens, and a T5 Transformer encoder to fuse image and text representations. Additionally, it has three separate predictors: a heatmap predictor for attention/saliency heatmaps or visual importance heatmaps, a scanpath predictor for the sequence/order of viewing, and a rating predictor for quality/aesthetic scores of images or web pages.

3. Methodology

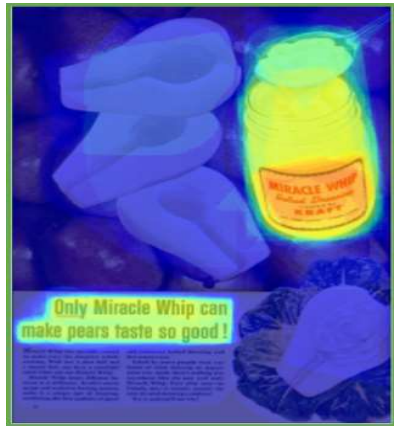
Heatmap Predictor



saliency heatmap



importance heatmap

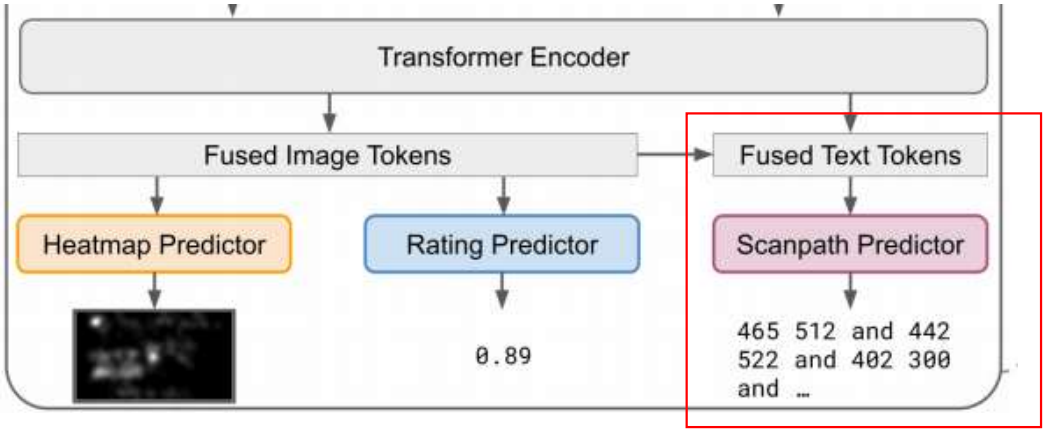


The heatmap prediction head takes the fused image tokens after the Transformer encoder, and processes the features via several read-out convolution layers, together with up-sampling so that the output will match the resolution of the input image. A sigmoid function is used at the end to ensure the generated values fall within the range [0, 1] for each pixel.

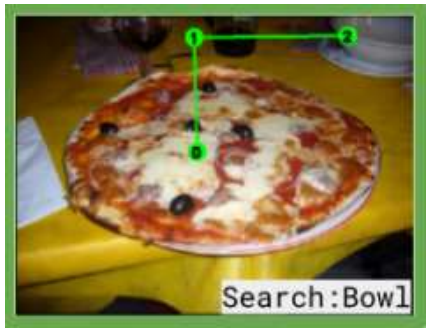
This paper adopt a pixel-wise $\ell 2$ loss function for the heatmap predictor during training.

3. Methodology

Scanpath (Sequence) Predictor



Search



Free-viewing

The scanpath predictor takes both the fused image and text tokens after the Transformer encoder as input, and applies a Transformer decoder to generate the predicted scanpath

A scanpath is defined as a sequence of 2D locations $(x_1, y_1),(x_2, y_2), \ldots,(x_N, y_N)$

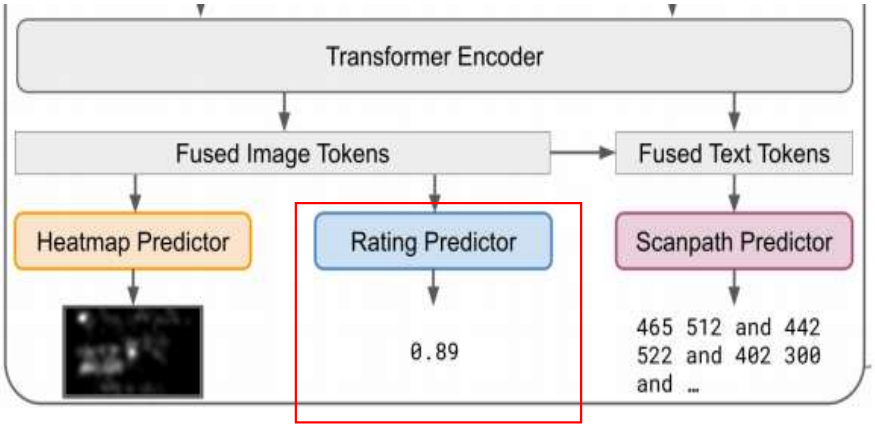
$$y = \texttt{<extra\_id\_01>}\ \tilde{x}_1\ \tilde{y}_1\ \texttt{and}\ \tilde{x}_2\ \tilde{y}_2\ \texttt{and}\ \ldots \hookrightarrow \texttt{and}\ \tilde{x}_N\ \tilde{y}_N\ \texttt{<extra\_id\_02>.}$$

$$\max \sum_{j=1}^{3N+1} w_j \log P(\tilde{y}_j | x, y_{1:j-1})$$

where x is the input image and text prompt, and y is the target sequence associated with x. wj is the weight for the j-th token that can adapt weights for different types of tokens. We use a unified weight for each token in experiments.

3. Methodology

Rating Predictor



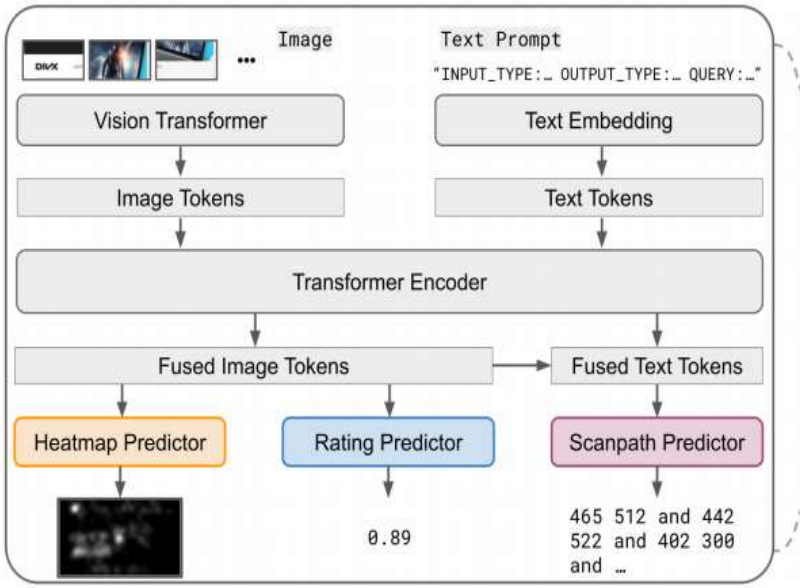
nature map

web map

This paper take a simple prediction head which takes image tokens after the Transformer encoder module, and processes the features via a few convolution and connected layers. An $\ell 2$ loss is used for training the rating predictor with rating data.

3. Methodology

Text Prompt




```
INPUT_TYPE:<input_type>
OUTPUT_TYPE:<output_type>
QUERY:<query>
```

We fill <input_type> with string taken from {natural image | web page | graphic design | mobile user interface} and <output_type> taken from {saliency heatmap | importance heatmap | aesthetics score | scanpath}. We append a query in string Query:<query> to the prompt if a task-specific query is available, for example, the object name to search, or the question to answer, depending on the use case. The prompt we use is modularized and can easily adapt to different types of datasets and scenarios.

4. Experiment

Datasets

Table 1. List of all public datasets used to train our model. '# Image' denotes the number of unique images in the entire dataset. Note that for annotation 'scanpath,' there are multiple scanpaths recorded from a group of users associated with one image, so '# Training Sample' is much larger than '# Image.' During training, we randomly sample from all training datasets with an equal sampling rate.

Dataset	Image domain	Training annotation	Viewing style	# Image	Image Resolution	# Training Sample
Salicon [34]	Natural scene	Saliency heatmap	Free-viewing	15,000	640 × 480	10,000
OSIE [65]	Natural scene	Saliency heatmap	Free-viewing	700	800 × 600	500
WS-Saliency [11]	Web page	Saliency heatmap	Free-viewing	450	1,280 × 720	392
Mobile UI [42]	Mobile user interface	Saliency heatmap	Free-viewing	193	Varied	154
Imp1k [25]	Graphic design	Importance heatmap	N/A	998	Varied	798
WS-Scanpath [11]	Web page	Scanpath	Free-viewing	450	1,280 × 720	5,448
FiWI [60]	Web page	Saliency	Free-viewing	159	1,360 × 768	121
COCO-Search18 [16]	Natural scene	Scanpath	Object-searching	3,101	1,680 × 1,050	21,622
COCO-FreeView [16]	Natural scene	Scanpath	Free-viewing	3,101	1,680 × 1,050	37,038
Koniq-10k [30]	Natural scene	Rating	N/A	10,073	1,280 × 720	7,000
Web Aesthetics [22]	Web page	Rating	N/A	398	1,280 × 768	398

4. Experiment results

Table 2. Heatmap prediction results on five public datasets spanning digital design and natural scene, benchmarking with eight metrics in total. The details on datasets and evaluation metrics can be found in Sec. 4.1. For *Imp1k* dataset we predict the importance heatmap, while for the rest of the four datasets, we predict attention/saliency heatmap. Our method is highlighted with green background . For each dataset and each metric, the best result in the current column is in **bold**, and the second best result is in blue. For our model, the relative performance change compared to the second best result (or the best result if we are not the best) in % is noted.

Dataset	Method	CC ↑	KLD ↓	AUC-Judd ↑	sAUC ↑	SIM ↑	NSS ↑	RMSE ↓	R^2 ↑
Mobile UI [42] (Mobile interface)	Itti et al. [32]	0.082	-	0.223	-	0.558	0.126	-	-
	BMS [73]	0.131	-	0.249	-	0.206	0.138	-	-
	GBVS [27]	0.580	-	0.666	-	0.709	0.591	-	-
	ResNet-Sal [42]	0.657	-	0.692	-	0.734	0.704	-	-
	SAM-S2015 [19]	0.477	-	0.650	-	0.562	0.537	-	-
	SAM-S2017 [19]	0.834	-	0.723	-	0.819	0.839	-	-
	SAM-mobile [42]	0.621	-	0.666	-	0.664	0.655	-	-
	Unified Model - Ours	**0.871** +4.44%	**0.120**+0.00%	**0.756** +4.56%	-	**0.826** +0.85%	**1.001** +19.31%	**0.121**+0.00%	**0.763**+0.00%
WS-Saliency [11] (Web page)	Itti et al. [32]	0.367	0.840	0.710	0.661	-	0.769		
	Deep Gaze II [39]	0.574	3.449	0.815	0.644	-	1.380	-	-
	SalGAN + *WS* [52]	0.637	0.622	0.818	0.703	-	1.458	-	-
	DVA [64]	0.571	0.701	0.805	0.711	-	1.260	-	-
	UAVDVSM [28]	0.519	0.858	0.739	0.668	-	1.133	-	-
	SAM-ResNet [19]	0.596	1.506	0.795	0.717	-	1.284	-	-
	EML-NET [33]	0.565	2.110	0.790	0.702	-	1.277	-	-
	UMSI [25]	0.444	1.335	0.757	0.698	-	1.042	-	-
	TaskWebSal-FreeView [76]	0.525	0.784	0.769	0.714	-	1.107	-	-
	SAM-ResNet + *WS* [19]	0.718	0.994	0.828	0.725	-	1.532	-	-
	DI Net + *WS* [67]	0.798	0.690	0.852	0.739	-	1.777	-	-
	AGD-F (W/o-L) [11]	0.815	0.637	0.858	0.753	-	**1.802**	-	-
	Unified Model - Ours	**0.836** +2.58%	**0.287** -54.95%	**0.863** +0.58%	**0.783** +3.98%	**0.744**+0.00%	1.792 -0.55%	**0.084**+0.00%	**0.705**+0.00%
FIWI [60] (Web page)	DeepGaze II [39]	0.488	-	0.797	0.625	-	1.229	-	-
	SAM-ResNet [19]	0.595	-	0.791	0.673	-	1.246	-	-
	UMSI [25]	0.457	-	0.755	0.675	-	0.938	-	-
	AGD-F [11]	0.735	-	0.767	0.748	-	1.606	-	-
	EML-NET [33]	0.661	0.603	0.847	0.675	-	1.653	-	-
	EML-NET + *Salicon* [33]	0.689	0.567	0.848	0.697	-	1.722	-	-
	Chen et al. [12]	0.699	0.564	0.851	0.704	-	1.752	-	-
	Unified Model - Ours	**0.740** +0.68%	**0.537** -4.79%	**0.861** +1.18%	**0.783** +4.68%	-	**1.855** +5.88%	-	-
Salicon [34] (Natural scene)	SalGAN [52]	0.763	-	-	**0.755**	-	-	-	-
	FastSal w. MobileNetV2 [31]	0.8751	-	0.845	0.736	-	1.816	-	-
	SimpleNet w. ResNet-50 [55]	0.895	0.211	0.868	-	0.786	1.881	-	-
	SimpleNet w. PNASNet-5 [55]	**0.907**	**0.193**	**0.871**	-	**0.797**	1.926	-	-
	MDNSal [55]	0.899	0.217	0.868	-	**0.797**	1.893	-	-
	Unified Model - Ours	0.901 -0.66%	0.215 +11.40%	0.870 -0.11%	0.752-0.40%	0.792 -0.63%	**1.947** +1.09%	**0.077**+0.00%	**0.813**+0.00%
OSIE [62] (Natural scene)	SALICON [34]	0.685	0.575	0.846	-	0.600	1.641	-	-
	SAM-ResNet [19]	0.758	**0.480**	0.860	-	0.648	1.811	-	-
	UMSI [25]	0.746	0.513	0.856	-	0.631	1.788	-	-
	EML-NET [33]	0.717	0.537	0.854	-	0.619	1.737	-	-
	Chen et al. [12]	**0.761**	0.506	0.860	-	**0.652**	1.840	-	-
	Unified Model - Ours	0.754 -0.92%	0.547 +13.96%	**0.867** +0.81%	**0.739**+0.00%	0.647 -0.46%	**1.842** +0.11%	**0.100**+0.00%	**0.575**+0.00%
Imp1k [25] (Graphic design)	Bylinskii et al. [8]	0.758	0.301	-	-	-	-	0.181	0.072
	Bylinskii et al. [8]	0.732	0.388	-	-	-	-	0.205	0.061
	SAM [19]	0.866	0.166	-	-	-	-	0.168	0.108
	UMSI-nc [25]	0.802	0.177	-	-	-	-	0.152	0.095
	UMSI-2stream [25]	0.852	0.168	-	-	-	-	0.141	0.105
	UMSI [25]	0.875	0.164	-	-	-	-	0.134	0.115
	Unified Model - Ours	**0.904** +3.31%	**0.123** -25.00%	-	-	**0.845**+0.00%	-	**0.079** -41.04%	**0.823** +615.65%

Table 4. Subjective rating prediction results on natural scene image dataset *KonIQ-10k* and web page dataset *Web Aesthetics*.

Dataset	Method	SRCC ↑	PLCC ↑
KonIQ-10k [30] (Natural scene)	BRISQUE [47]	0.665	0.681
	ILNIQE [74]	0.507	0.523
	HOSA [66]	0.671	0.694
	BIECON [37]	0.618	0.651
	WaDIQaM [7]	0.797	0.805
	PQR [72]	0.880	0.884
	SFA [44]	0.856	0.872
	DBCNN [75]	0.875	0.884
	MetaIQA [77]	0.850	0.887
	BIQA (25 crops) [61]	**0.906**	0.917
	MUSIQ-single [36]	0.905	0.919
	Ours	0.905 -0.11%	**0.925** +0.65%
Web Aesthetics [22] (Web page)	Rating-based Calista [22]	-	0.770
	Comparison-based Calista [22]	-	0.820
	Ours	**0.813** +0.00%	**0.841** +2.56%

4. Experiment results

Table 3. Scanpath prediction results on natural scene and digital design datasets, with object-searching and free-viewing tasks.

Dataset	Method	SemSS ↑	SemFED ↓	Sequence Score ↑	Shape ↑	Direction ↑	Length ↑	Position ↑	MultiMatch ↑
COCO-Search18 [16] (Natural scene, object searching)	IRL [68]	0.481	2.259	-	0.901	0.642	0.888	0.802	0.833
	Chen et al. [14]	0.470	1.898	-	0.903	0.591	0.891	0.865	0.820
	FFM [69]	0.407	2.425	-	0.896	0.615	0.893	0.850	0.808
	Gazeformer [48]	0.496	**1.861**	-	0.905	**0.721**	0.857	**0.914**	0.849
	Unified Model - Ours	**0.521** +5.04%	2.004 +7.68%	-	**0.946** +4.53%	**0.724** +0.42%	**0.924** +3.47%	0.901 -1.42%	**0.874** +2.94%
WS Scanpath [11] (Web page, free-viewing)	Itti et al. [32]	-	-	0.177	0.781	0.676	0.778	0.594	0.707
	MASC [2]	-	-	0.169	0.788	0.580	0.818	0.514	0.717
	SceneWalker [59]	-	-	0.194	0.843	0.616	0.842	0.562	0.716
	G-Eymol [71]	-	-	0.218	0.820	0.673	0.816	0.681	0.748
	AGD-F (w. layout) [11]	-	-	0.203	0.787	0.642	0.771	0.677	0.719
	AGD-S (w/o layout) [11]	-	-	0.221	0.814	0.663	0.805	0.698	0.745
	AGD-S (w. layout) [11]	-	-	0.224	0.820	0.677	0.813	0.708	0.755
	Unified Model - Ours	-	-	**0.267** +19.20%	**0.967** +14.71%	**0.826** +22.01%	**0.960** +14.01%	**0.794** +12.15%	**0.887** +17.48%

5. Ablation

Transferring Knowledge between Tasks

In this section, we demonstrate the zero-shot generalization ability between image domains and tasks of UniAR. Here, zero-shot generalization refers to a model's ability to perform on completely unseen image domain and prediction tasks without requiring dedicated training.

Table 5. Experiment on zero-shot generalization to *WS-Scanpath* dataset, described in Sec. 4.4. *CC* = *COCO-FreeVeiw* dataset.

Training Set	Sequence Score ↑	MultiMatch ↑
WS-Scanpath (SOTA results from [11])	0.224	0.755
WS-Scanpath (ours)	0.261	0.894
UniAR full model	0.267	0.887
CC scanpath (ours)	0.196	0.836
CC scanpath + *WS-Saliency* (ours)	0.190	0.858
CC saliency/scanpath + *WS-Saliency* (ours)	0.231	0.857

Our experiment uses WS (web page) and COCO Freeview (natural scene) datasets. We assess our model on scanpath prediction on the WS-Scanpath dataset. We vary the training sets in three different scenarios:
(1)Using scanpath data from COCO-Freeview;
(2) Combining scanpath data from COCO-Freeview with saliency heatmaps from WS;
(3) Employing both scanpath and saliency heatmap data from COCO-Freeview, augmented with saliency heatmap data from WS.

Thanks !