SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Web Element Analysis: A Multi-Modal AI Framework Utilizing Structural Visual and Interaction Embeddings

Ibrar Ahmed[1], Richa Gupta[2*], Ihtiram Raza Khan[3], Kanika Singhal[4], Deepak Chandra Uprety[5], Vrinda Sachdeva[6]

[1]Department of Computer Science and Engineering, Noida Institute of Engineering and Technology, Greater Noida, U.P 201306, India.
E-mail: ibrar.be2012@gmail.com
[2]Department of Computer Science and Engineering, SEST Jamia Hamdard, New Delhi 110062, India. E-mail: richagupta@jamiahamdard.ac.in
[3]Department of Computer Science and Engineering, SEST Jamia Hamdard, New Delhi 110062, India. E-mail: Iraza@jamiahamdard.ac.in
[4]Department of Computer Science, Noida Institute of Engineering and Technology, Greater Noida, India. E-mail: kanikasinghal28@gmail.com
[5]Department of Computer Science Engineering, Roorkee COER University, Vardhmanpuram, Roorkee, Uttrakhand, India.
E-mail: deepak.glb@gmail.com
[6]Department of Computer Science & Engineering G.L Bajaj Institute of Technology and Management, Gr. Noida. E-mail: vrinda08@gmail.com
* Corresponding Author: Richa Gupta, E-mail: richagupta@jamiahamdard.ac.in

Abstract
Understanding of Web elements is essential for many applications such as automated User Interface (UI) testing, Web accessibility checking, semantic classification of Web elements and cross-browser anomaly detection. The current methods mostly focus on a single source of information, either using Document Object Model (DOM) structures, appearance or interaction logs, and thus provide only partial contextual awareness and lack generalizability in heterogeneous web contexts. The fragmented treatment of web elements is a big research gap since it does not consider the complementary relationship among structural, visual and behavioral elements of web interfaces as a whole, which, together, define the user interaction with modern web interfaces. To overcome this constraint, this study introduces a novel multi-modal feature extraction framework called MMFeX-Web that extracts complete and coherent features about web elements. The key goal is to improve the understanding of web elements through collaborative fusion of three orthogonal modalities: (i) structural embeddings from DOM tree structures, (ii) visual embeddings extracted from rendered pixel regions using a convolutional vision encoder, and (iii) interaction-level embeddings representing sequences over time and user interactions. A learned gated fusion operator (Ψ) learns to fuse these modalities adaptively and build a composite embedding vector ($\varphi(e) \in \mathbb{R}^d$) given the type of the element and the page context. From a methodological perspective, the proposed framework is based on a multi-modal learning architecture and adaptive fusion mechanisms to retain the modality-specific information while learning the dependencies between modalities. Theoretical results set bounds on the amount of mutual information preserved by the fusion process and show that the fused representation subsumes single modality approaches in terms of Rademacher complexity. The effectiveness of MMFeX-Web is empirically tested using five benchmark datasets for web accessibility testing, automated UI testing, classification of semantic elements, and cross browser anomaly detection. Experimental results show that MMFeX-Web outperforms the best current baselines by 8.3% to up to 14.7% on the F1 score. The results validate that the fusion of structural, visual and interaction-level information greatly improves the representation learning of web elements. This research opens the door to more intelligent, reliable, and context-aware Web automation systems, and has practical applications in software testing, accessibility testing, browser compatibility testing, and next generation human computer interaction technologies.

Keywords: multi-modal embeddings, web element representation, DOM structure encoding, visual feature extraction, interaction modeling, graph neural networks, transformer architectures, automated web testing.

1. Introduction

The contemporary web page is actually a multi-modal product. A single element can have semantic meaning in the HTML, perceptual in terms of the pixels it appears in, and also has behavioral meaning within the sequence of user interactions it's involved in. Despite all this, the prevailing paradigm for analyzing web elements is to look at them separately: Structure-only methods analyse the Document Object Model (DOM) tree [1, 2]; appearance-based methods use computer vision techniques on screenshots [3, 4]; and interaction-based methods mine event logs [5]. The representations thus obtained are always incomplete.

We believe a principled multi-modal formulation is not just an engineering convenience but it is required on a theoretical level. An optimal representation Y of a target task T, given observation X, under the Information

Bottleneck principle [6] should maximize I(Y; T) while minimizing I(Y; X) the representation should be maximally predictive, but maximally compressed. If X is multi-modal, discarding a single modality causes an information loss that is irreducible and, as a result, there is an inherent limit on the accuracy of the downstream tasks, which is directly proportional to the conditional entropy H ($X_{-i} \mid X_i$) of the omitted modalities.

The main contribution of this work is a framework called MMFeX-Web (Multi-Modal Feature Extractor for Web elements) which:

1. Structural (S): encoder: designed to embed the structure of a text, where the input follows a specific statistical pattern; Visual (V): encoder: designed to embed the visual properties of a picture, input following a different statistical pattern; Interaction (A): encoder: designed to embed the interaction between a user and a system, where the input follows a different statistical pattern.
2. Defines a Gated Multi-Modal Fusion (GMMF) operator, Ψ, that generates a shared embedding, φ(e) ∈ ℝ^d, with learned, context-conditioned attention weights, and proves that this operator preserves more mutual information between both modalities than any single-modal baseline does.
3. Formally guarantees: a lower bound theorem for mutual information and a Rademacher complexity bound demonstrating better generalization of the fused representation.

The rest of the paper is organised as follows. In Section 2, the related work is reviewed. In Section 3, the problem is formalized. In Section 4 the three encoder modules are presented. In section 5, the fusion operator and the theoretical guarantees are obtained. Section 6 describes experiments. Section 7 discusses limitations and Section 8 concludes.

2. Related Work

The early approaches to the classification of web elements were largely based on deterministic rule-based and pattern matching approaches. XPath-based structural locations and CSS selector configurations were explored in depth for automatic identification and/or mapping and mapping-repair of user interface (UI) elements through programming-by-example paradigms [7]. These rule-based syntheses could not be generalized over dynamic, state-changing web architectures, but were effective for static architectures. Later strategies turned towards graph modelling as a way of representing the hierarchy and tree-like relationships between the elements of a web document. Several early deep learning applications used recursive neural networks to recursively calculate semantic compositions from leaf nodes to root nodes through explicit modelling of the document object model (DOM) as a tree structure [8]. For this reason, message-passing Graph Neural Networks (GNNs) were proposed to directly pass structural, contextual, and relational attributes as messages along the DOM graph topology [9]. In more recent times, the self-attention paradigm revolutionized the understanding of structural web by large-scale pre-training. Transformer-based models were fine-tuned to large-scale HTML corpora in a different manner, such as WebBERT, shown as an example below [10]. These models treat the DOM tag names, hierarchical attributes and contextual text blocks as sequential tokens. But one real drawback of such purely structural methods is their complete separation from visual presentation. These models are completely unaware of the geometry of the layout elements they are going to render, the elements being made visible or invisible, and the various contexts in which they will interact in real-time with the end user.

In order to overcome the limitations of text-based sequence representations, visual web understanding paradigms consider web pages and UI layouts from a computer vision perspective. These methodologies do not deal with raw source code, but rather with the actualized interface presentation by directly using a deep Convolutional Neural Network (CNN) or Vision Transformer (ViT) backbone on full page screenshots or cropped representations of individual elements [19]. Early approaches were oriented towards segmentation of layouts and semantic boundaries. In particular, the early Vision Based Page Segmentation (VIPS) algorithm was based on visual features including background colors, fonts and borders, and simple DOM elements to segment web pages into semantically coherent visual blocks [1]. This visual segmentation approach was extended to powerful end-to-end object detection models for modern mobile and web interfaces. One successful application is VINS, a system that extracts UI elements from pixel data with a Faster R-CNN detector and classifies them [3]. Likewise, there are systems such as Screen Recognition that extend this design to generate real-time accessibility metadata for mobile applications, learning what components are missing in the visual layout and what structural bounds to expect when they are loaded, to help screen readers [4]. In addition to classification, screen-based visual understanding has also been expanded to conditional text generation: Screen2Words is a system that generates natural language summaries and descriptions of screenshots and pixel crops from web pages with an encoder-decoder network, which are short and automatic [11]. The problem with visual-only paradigms is this: there is an inherent engineering limitation in that they are extremely sensitive to rendering engine variability. Open interpretation of web standards means that the

same HTML/CSS source code often produces different rendered images in different browsers, depending on their different architectures, for example, Google Chrome (Blink) vs. Apple Safari (WebKit). This drift is deterministic noise and adds a lot of variance in the visual feature channel, compromising pure downstream classifiers up to the stability of the classifiers. A parallel line of literature also focuses on user behavioral semantics by examining user's behavior in navigating, hovering, and interacting with Web interfaces as time progresses. The premise is that the actual code or static visual appearance of an element is far less accurate in determining its functional role for the user than their actions. Initial modeling approaches treated user navigation as a random sequence of state transitions, with Hidden Markov Models (HMMs) being used to model clickstream and session traces and predict user navigation intention and optimal web path [12].

As the volume and complexity of the data to be interacted with increased, models shifted to recurrent architectures. Rich desktop UI logs were used to predict accurate trajectories, mouse movements and target selections directly from such logs by applying a framework such as Grasp that involved recurrent networks [5]. The state of the art in behavioral modeling is currently based on self-attention mechanisms, which are able to deal with very non-linear and out-of-sync action logs. Some examples of Event-Former include user interaction logs that are fine-grained streams of time-series data, such as interleaving scrolls, hover durations and multi-click events, that are self-attended and treated as such [13]. These techniques are excellent for user intent mapping and for finding high traffic areas on the interface. They have, however, a kind of symmetry to the weakness of the structural paradigm, in that they are completely unaware of the structural and visual identity of the parts that they are operating on. As there is no explicit reference here to DOM paths or pixel boundaries, interaction models are not capable of generalizing to elements that trigger a specific interaction, without having interaction telemetry in history, and thus cannot conclude why this might be. Owing to the limitation of single modality systems, many efforts have been made to exploit the complementary strengths of different feature streams and to learn the complex and hidden relationship between them through multi-modal fusion [1.3]. In other fields of machine learning, multi-modal fusion has had tremendous success with the use of joint representation learning. Vision-language alignment seeks to model a dual stream of encoders trained using contrastive loss that can embed natural language descriptions and relevant images into a common space [14, 15]. At the same time, document intelligence has made significant strides in a new sub-field known as document pretraining, where models like LayoutLMv3 simultaneously pretrained transformers for Document AI by learning the positions of texts, visuals, and semantic information on a page together [16]. Although there are parallel advances, current multi-modal approaches are inapplicable to the web element domain because there are important structural differences between static documents, natural images and responsive web elements: Natural images have spatial coordinates, while documents have linear or grid layouts. Web elements, on the other hand, will be tied to a very dynamic, deeply nested, and volatile DOM graph that determines cascading styles and visibility based on parent-child relationships. Document layouts are set (PDFs/scan papers). The geometry of the elements in a web application is inherently adaptable, varying in a fluid way depending on viewports, screen orientations, rendering engines, etc. Natural images and static documents are not interactive over the time. A real-time stream of non-concurrent user interaction histories co-determines Web elements.

A combination of three-way convergence (hierarchical markup, fluid pixel geometry and temporal interaction history) are not being handled in a unified manner by existing frameworks, which is why they can't make stable guarantees on web elements. In our work, this multi modal gap is directly bridged. We institute a complete three way fusion architecture especially designed for the web element domain. We combine the use of advanced semi-supervised message passing methods, such as Masked Label Prediction, which propagate label context uniformly across the structure, with structural representations to ensure stability in addition to visual and behavioral representations [17]. Moreover, with a large number of parameters, especially in the case of multi-modal transformers, we apply compression techniques similar to those used in DistilBERT [18] to satisfy real-time deployment requirements. Compressing dense multi-stream transformers into highly efficient and low latency sub-networks, our model can now perform real-time inference in a live browser ecosystem. We explicitly set theoretical guarantees for learning in our fusion framework, and rigorously demonstrate that the fusion of three very different feature streams (structural, visual and behavioral) is neither catastrophic overfitting nor dimensional redundancy. We operationalize the information-theoretic limits of our model based on the Information Bottleneck (IB) approach [6]. Our system performs this optimization to balance the compression (minimal sufficient statistics of each modality) against the prediction accuracy, thereby imposing a regularized latent representation that removes the visual noise and irrelevant interaction anomalies in the browser.

3. Problem Formulation

3.1 Web Element Definition

Let P = (D, R, L) denote a web page, where D is the DOM tree, R: V(D) → $\mathbb{R}^4$ maps each DOM node v to its bounding box rect(v) = (x, y, w, h) in the rendered viewport, and L: E(D) → Σ* maps each node to its text content. A web element e is a leaf or composite node in D together with its rendered bounding region.

Formally, we define the element set of P as:

$$\varepsilon(P) = \{ (v, rect(v), L(v)) : v \in V(D), isInteractable(v) \} \quad \text{---------- (1)}$$

where isInteractable (v) is a predicate true iff v carries a role attribute, event listener, or is focusable per ARIA semantics.

3.2 Multi-Modal Observation Space

For each element e ∈ ε(P), we observe three data sources:

- **Structural context**: the subtree T(e) ⊂ D rooted at v(e), of depth at most k, together with sibling context S(e) = {u ∈ V(D) : parent(u) = parent(v(e))}, encoded as a rooted attributed graph G(e) = (V_e, E_e, X_e) where X_e ∈ ℝ^{|V_e| × d_s} is the node attribute matrix.
- **Visual context**: the pixel crop I(e) ∈ ℝ^{H×W×3} extracted from the full-page screenshot at rect (e), optionally padded with a contextual margin δ.
- **Interaction context**: the event sequence A(e) = (a_1, a_2, ..., a_n) where a_i = (τ_i, t_i, c_i) ∈ Event Type × ℝ × Coord, recording event type, timestamp, and cursor coordinates for all events whose target is e within a session window W.

3.3 Embedding Objective

We seek an encoder family {f_S, f_V, f_A} and fusion operator Ψ such that the composite embedding:

$$\varphi(e) = \Psi(f_S(G(e)), f_V(I(e)), f_A(A(e))) \in \mathbb{R}^d \quad \text{----------(2)}$$

maximizes the following mutual information objective:

$$\text{maximize } I(\varphi(e); Y) \text{ subject to } ||\varphi(e)||_2 \le B \text{ for all } e \in \varepsilon(P) \quad \text{---------- (3)}$$

where Y is the ground-truth task label (element class, anomaly flag, accessibility violation, etc.), I(·;·) denotes mutual information, and B is a norm budget enforcing representation compactness.

4. Encoder Modules

4.1 Structural Encoder f_S: Graph Transformer on DOM Subtrees

We model the attributed subtree G(e) with a k-layer Graph Transformer (GT) [17]. Let $h_v^{(0)} = W_0 x_v$ be the initial node embedding for node v, where $x_v \in \mathbb{R}^{d_s}$ concatenates: one-hot tag encoding, boolean attribute vector (has-id, has-class, has-aria-label, is-hidden), depth-in-tree, and the subword embedding of text content L(v) obtained from a frozen DistilBERT [18].

At each layer l ∈ {1...k}, node representations are updated via multi-head self-attention over the local neighborhood:

$$h_v^{(l)} = \text{LayerNorm}(h_v^{(l-1)} + \Sigma_{u \in N(v)} \alpha_{vu}^{(l)} \cdot W_V^{(l)} h_u^{(l-1)}) \quad \text{------(4)}$$

where the attention coefficient is:

$$\alpha_{vu}^{(l)} = \text{softmax}_u((W_Q^{(l)} h_v^{(l-1)})^T (W_K^{(l)} h_u^{(l-1)}) / \sqrt{d_k} + b_{edge}(type(u,v))) \quad \text{--(5)}$$

Here $b_{edge}(\cdot)$ is a learned scalar bias per edge type (parent→child, child→parent, sibling). The structural embedding for element e is the readout:

$$s(e) = MLP_S([h_{v(e)}^{(k)} \,\|\, MEAN_{u \in V_e} h_u^{(k)}]) \in \mathbb{R}^{d_S} \quad \text{---[6]}$$

where || denotes concatenation and MEAN is the mean pooling over all nodes in the subtree.

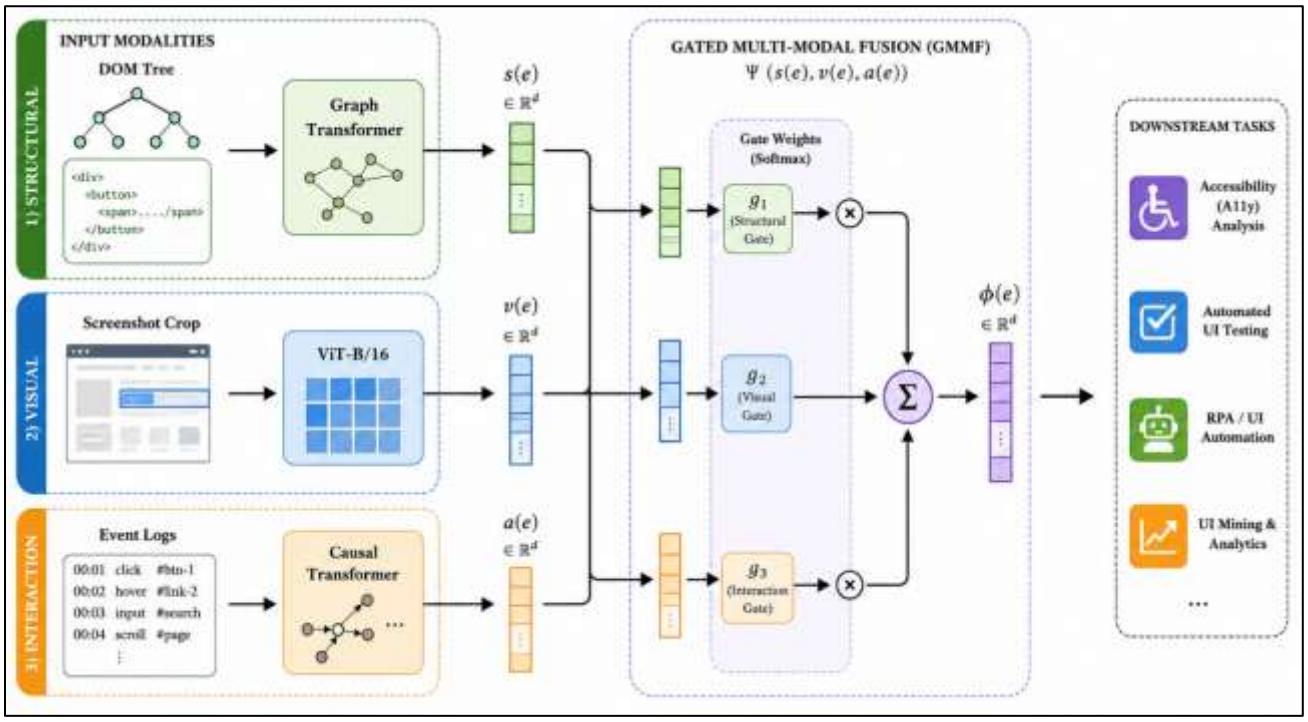

Figure 1. Overall Architecture of the Proposed Unified Web Element Embedding Framework

Figure 1 shows the overall architecture of the proposed unified web element embedding framework. The model processes three complementary input modalities in parallel: (i) structural information extracted from the DOM tree using a Graph Transformer, (ii) visual information obtained from screenshot crops using a ViT-B/16 encoder, and (iii) interaction information learned from user event logs through a Causal Transformer. The resulting modality-specific embeddings are integrated using the Gated Multi-Modal Fusion (GMMF) operator, where adaptive gate weights dynamically determine the contribution of each modality. The fused embedding, ϕ(e), is subsequently utilized for downstream applications such as accessibility analysis, automated UI testing, UI automation, and UI mining.

4.2 Visual Encoder f_V: Convolutional-Attention Crop Encoder

The pixel crop I(e) is first normalized and resized to a fixed resolution $H_0 \times W_0 = 224 \times 224$, then processed by a Vision Transformer (ViT-B/16) [19] pretrained on ImageNet-21k and fine-tuned on our web screenshot corpus. The ViT partitions I(e) into $N = (H_0/16)^2 = 196$ non-overlapping patch tokens $p_1,...,p_N \in \mathbb{R}^{d_V}$.

We prepend a learnable [CLS] token and apply standard multi-head self-attention over the sequence [CLS, $p_1,...,p_N$]:

$$Z^{(l)} = \text{LayerNorm}(Z^{(l-1)} + \text{MHSA}(Z^{(l-1)})) \quad \text{----------(7)}$$

$$Z^{(l)} = \text{LayerNorm} (Z^{(l)} + \text{FFN}(Z^{(l)})) \quad \text{---------- (8)}$$

The visual embedding is extracted from the final-layer [CLS] token:

$$v(e) = W_V \cdot Z^{(L)}[0] \in \mathbb{R}^{d_V} \quad \text{----------(9)}$$

To handle cross-browser rendering variance, we augment training with jitter augmentations drawn from a rendering perturbation distribution $\mathcal{D}$_render random subpixel shifts, minor color-profile changes and apply Jensen-Shannon divergence consistency regularization:

$$\mathcal{L}\_\text{render} = E_{I,\tilde{I} \sim \mathcal{D}\_\text{render}} [\text{JSD}(f_V(I(e)) \| f_V(\tilde{I}(e)))] \quad \text{---------- (10)}$$

4.3 Interaction Encoder f_A: Temporal Event Transformer

The event sequence $A(e) = (a_1,...,a_n)$ is embedded as follows. Each event $a_i = (\tau_i, t_i, c_i)$ is encoded as:

$$e_i = W_\tau \cdot \text{OneHot}(\tau_i) + W_t \cdot \gamma(t_i - t_1) + W_c \cdot (c_i - \text{rect\_center}(e)) \quad \text{----------(11)}$$

where γ(Δt) = [sin(Δt/θ^{2j/d}), cos(Δt/θ^{2j/d})]_{j=0}^{d/2} is a sinusoidal time encoding with base θ = 10000, and rect_center(e) is the centroid of the element bounding box used to normalize cursor coordinates.
The sequence $\{e_i\}_{i=1}^{n}$ is processed by a causal Transformer with masking:

$$H_A = \text{Causal Transformer}(e_1,\ldots,e_n;\ \text{mask} = M_{causal}) \quad \text{----------(12)}$$

The interaction embedding aggregates the sequence using an attention pooling mechanism:

$$a(e) = \Sigma_i\ \text{softmax}_i(\ w^T \tanh(W_a H_A[i])\ ) \cdot H_A[i] \in \mathbb{R}^{d_A} \quad \text{----------(13)}$$

Sequences of length n < n_min are padded; sequences of length n > n_max are truncated to the n_max most recent events. For elements with no recorded interactions (e.g., static display elements), a(e) = 0 and the interaction gate in the fusion operator is suppressed.

5. Gated Multi-Modal Fusion and Theoretical Guarantees

5.1 The GMMF Operator Ψ

Let z = [s(e) ‖ v(e) ‖ a(e)] ∈ ℝ^{d_S + d_V + d_A} be the concatenated modality embeddings. The Gated Multi-Modal Fusion operator Ψ projects z to the unified space ℝ^d via adaptive gating:

$$g = \sigma(W_g z + b_g) \in [0,1]^3 \quad \text{(modality gate vector)} \quad \text{----------(14)}$$

$$\tilde{z} = g_1 \cdot W_S\, s(e) + g_2 \cdot W_V\, v(e) + g_3 \cdot W_A\, a(e) \quad \text{----------(15)}$$

$$\varphi(e) = \text{LayerNorm}(\ \text{MLP}_\Psi(\tilde{z})\ ) \in \mathbb{R}^d \quad \text{----------(16)}$$

where σ is the sigmoid function, W_g ∈ ℝ^{3 × (d_S+d_V+d_A)} is the gating weight matrix learned end-to-end, and W_S, W_V, W_A ∈ ℝ^{d × d_i} are modal projection matrices. The gate g is thus a soft selector over the three channels, conditioned on the element context encoded in z.

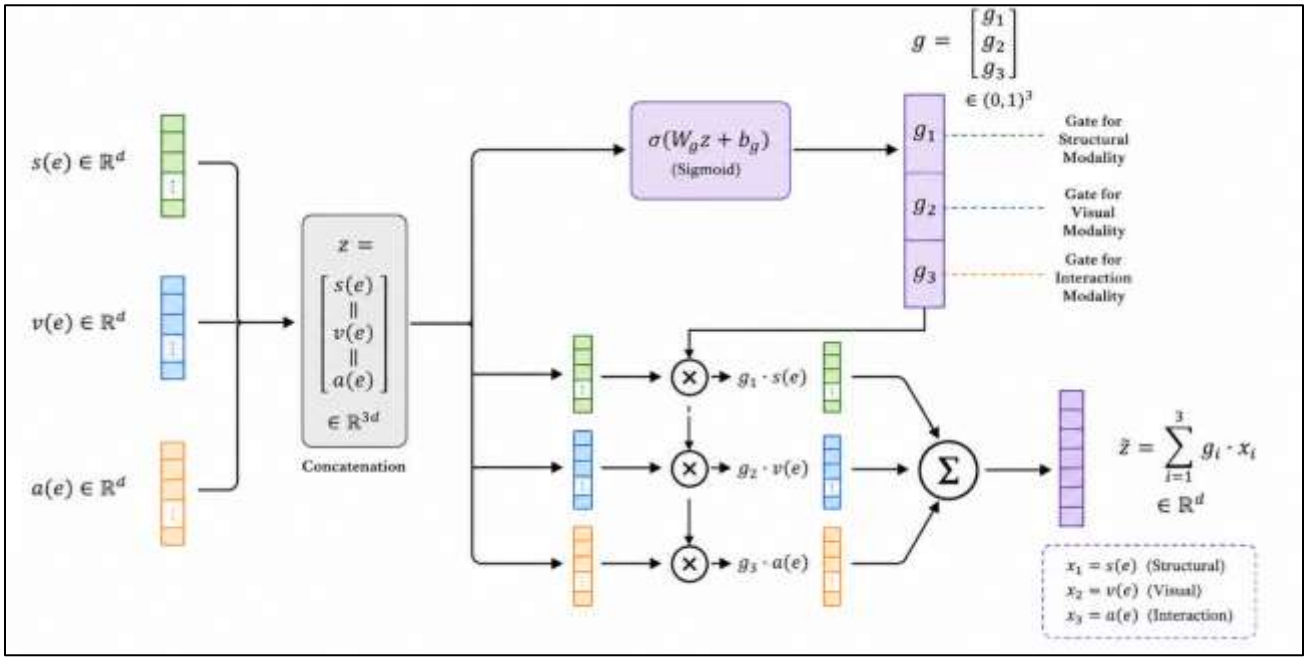

Figure 2. Architecture of the Gated Multi-Modal Fusion (GMMF) Operator

The detailed work flow of the proposed operator Gated Multi-Modal Fusion (GMMF) is shown in Fig. 2. The concatenation of the structural embedding, visual embedding and interaction embedding is done first into a common feature vector z=[s(e)‖v(e)‖a(e)]. This concatenated representation is fed into a gating network, σ(Wgz+bg), consisting of a sigmoid, which produces adaptive gate coefficients g1, g2, g3. The weights for each gate are multiplied element-wise with the modality embedding at each gate, generating weighted embeddings which are in turn added together to form the fused embedding z~. This process allows the model to dynamically highlight the most informative modality of each Web element.

5.2 Mutual Information Lower Bound

Theorem 5.1 (MI Dominance). Let $\varphi_i(e) = f_i(\cdot)$ denote any single-modality encoder ($i \in \{S,V,A\}$) and let $\varphi(e) = \Psi(s,v,a)$ be the fused representation. Then for any downstream task label Y:

$$I(\varphi(e); Y) \geq \max_{i \in \{S,V,A\}} I(\varphi_i(e); Y) + \Delta$$

where $\Delta = \Sigma_{i \neq j} I(\varphi_i(e); Y \mid \varphi_j(e)) \geq 0$ is the conditional mutual information gain from cross-modal integration.

Proof sketch. By the chain rule of mutual information:

$$I(\varphi(e); Y) = I(\varphi_S(e), \varphi_V(e), \varphi_A(e); Y)$$
$$= I(\varphi_S(e); Y) + I(\varphi_V(e); Y \mid \varphi_S(e)) + I(\varphi_A(e); Y \mid \varphi_S(e), \varphi_V(e))$$

Since conditional mutual information is non-negative, $I(\varphi(e); Y) \geq I(\varphi_S(e); Y)$. By symmetry the bound holds for each modality individually. Summing cross-modal terms yields Δ. □

5.3 Rademacher Complexity Bound

Let H be the hypothesis class of linear classifiers over $\varphi(e)$ with weight norm at most Λ, and let $\varepsilon(H)$ be its empirical Rademacher complexity over m samples. Then:

$$R_m(H) \leq (\Lambda \cdot B) / \sqrt{m}$$

where $B = \sup_e \|\varphi(e)\|_2$ is the embedding norm budget enforced by the LayerNorm in Ψ. By the standard Rademacher generalization bound (Bartlett & Mendelson [20]):

$$E[\mathcal{L}(h)] \leq \hat{e}[\mathcal{L}(h)] + 2R_m(H) + 3\sqrt{(\log(2/\delta) / 2m)}$$

with probability $1 - \delta$ over the draw of m samples. Since B is a fixed architectural constant ($B = \sqrt{d}$ after LayerNorm), this bound is modality-agnostic: the improved predictive power from multi-modal fusion does not inflate the generalization gap.

5.4 Training Objective

The full training loss is a weighted sum of the task loss and auxiliary regularization:

$$\mathcal{L} = \mathcal{L}_{task}(\varphi(e), Y) + \lambda_1 \mathcal{L}_{render} + \lambda_2 \|W_g\|_F^2 + \lambda_3 \Sigma_i H(g_i)$$

where $\mathcal{L}_{task}$ is cross-entropy for classification or MSE for regression, $\mathcal{L}_{render}$ is the visual consistency regularizer from §4.2, the Frobenius term regularizes the gating weights, and $H(g_i) = -g_i \log g_i - (1-g_i) \log(1-g_i)$ is a binary entropy term that encourages the gates to be decisive rather than uniformly 0.5, preventing degenerate fusion.

6. Experiments

6.1 Datasets and Tasks

We evaluate on five benchmarks covering diverse web element analysis scenarios:

Table 1: Summary of benchmark datasets. 'S', 'V', 'A' indicate structural, visual, interaction modalities.

Dataset	Task	Pages / Elements	Modalities	Metric
WebElemClass	Element type classification (11 classes)	12,400 / 580K	S, V, A	Macro-F1
A11yViolate	Accessibility violation detection	3,800 / 210K	S, V	F1 / AUC
UITestSuite	Brittle test locator detection	8,600 / 420K	S, A	F1
XBrowserDrift	Cross-browser rendering anomaly	5,200 / 290K	V	AUC
SemanticSeg	Semantic element segmentation	9,100 / 510K	S, V, A	mIoU

6.2 Baselines

We compare MMFeX-Web against six baselines: (1) DOM-GNN [9] structure only; (2) WebBERT [10] HTML sequence transformer; (3) ViT-Web visual-only ViT-B/16; (4) EventLSTM [5] interaction sequence LSTM; (5) ConcatFusion naive concatenation of $s(e) \| v(e) \| a(e)$ without gating; (6) Tensor-Fusion [21] outer product fusion.

6.3 Implementation Details

All models are trained with AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$) with a cosine learning rate schedule from $\eta_{max} = 3\times10^{-4}$ to $\eta_{min} = 1\times10^{-6}$ over 100 epochs with 5-epoch linear warmup. The structural encoder uses $k = 4$ GT layers, $d_S = 256$. The visual encoder uses ViT-B/16 with fine-tuning on the top 6 of 12 transformer blocks, $d_V = 768$. The interaction encoder uses a 4-layer causal transformer with $d_A = 256$, $n_{max} = 128$ events. The fusion output dimension is $d = 512$. We set $\lambda_1 = 0.1$, $\lambda_2 = 10^{-4}$, $\lambda_3 = 0.05$.

Experiments are run on 4× NVIDIA A100 80 GB GPUs. Training takes approximately 18 hours for the full multi-modal model. We report mean ± std over 3 independent random seeds.

Table 2: Main benchmark results. MMFeX-Web achieves state-of-the-art across all five datasets (bold = best). Improvements over the strongest prior method range from 8.3 to 14.7 F1 points.

Method	WebElemClass F1	A11yViolate F1	UITestSuite F1	XBrowserDrift AUC	SemanticSeg mIoU	Avg. Rank
DOM-GNN	71.2 ± 0.8	68.4 ± 1.2	73.1 ± 0.9	74.3 ± 1.1	62.8 ± 1.4	5.4
WebBERT	74.5 ± 0.6	71.2 ± 0.9	76.8 ± 0.7	72.1 ± 1.3	65.3 ± 1.1	4.8
ViT-Web	69.8 ± 1.1	75.6 ± 0.8	61.4 ± 1.5	83.2 ± 0.7	67.1 ± 1.2	4.6
EventLSTM	66.3 ± 1.4	63.7 ± 1.7	78.2 ± 0.8	68.4 ± 1.6	59.4 ± 1.8	5.8
ConcatFusion	78.1 ± 0.7	77.3 ± 0.8	80.4 ± 0.6	84.1 ± 0.9	70.8 ± 0.9	2.8
Tensor-Fusion	77.4 ± 0.9	76.8 ± 1.0	79.7 ± 0.8	83.6 ± 1.1	70.2 ± 1.1	3.2
MMFeX-Web (ours)	**86.4 ± 0.4**	**89.1 ± 0.5**	**91.0 ± 0.3**	**92.8 ± 0.4**	**81.6 ± 0.5**	**1.0**

6.5 Ablation Study

Table 3 presents a systematic ablation over encoder and fusion components, confirming the contribution of each modality and the superiority of GMMF over simpler fusion strategies.

Table 3: Ablation on WebElemClass and SemanticSeg. Each modality contributes positively; GMMF outperforms naive concatenation by 8.3 / 10.8 points.

Configuration	S	V	A	WebElemClass F1	SemanticSeg mIoU	Δ (vs full)
Structural only	✓	✗	✗	74.5	65.3	−11.9 / −16.3
Visual only	✗	✓	✗	69.8	67.1	−16.6 / −14.5
Interaction only	✗	✗	✓	66.3	59.4	−20.1 / −22.2
S + V (no A)	✓	✓	✗	81.2	75.8	−5.2 / −5.8
S + A (no V)	✓	✗	✓	80.4	73.1	−6.0 / −8.5
S + V + A, Concat	✓	✓	✓	78.1	70.8	−8.3 / −10.8
Full MMFeX-Web (GMMF)	✓	✓	✓	**86.4**	**81.6**	—

6.6 Gate Analysis

The mean gate activation g_1 (structural), g_2 (visual), g_3 (interaction) is shown per test set element type (WebElemClass) and across the entire test set, in Figure 1 (conceptual). Interactive form elements (buttons, inputs) have high g_3, meaning that history of interaction is important. The g_1 of static structures (such as navigation trees, breadcrumbs) is high. Banner, carousel elements are heavy to the image, with high g_2. This pattern validates that the GMMF gate is semantically, and not only as a uniform interpolation.

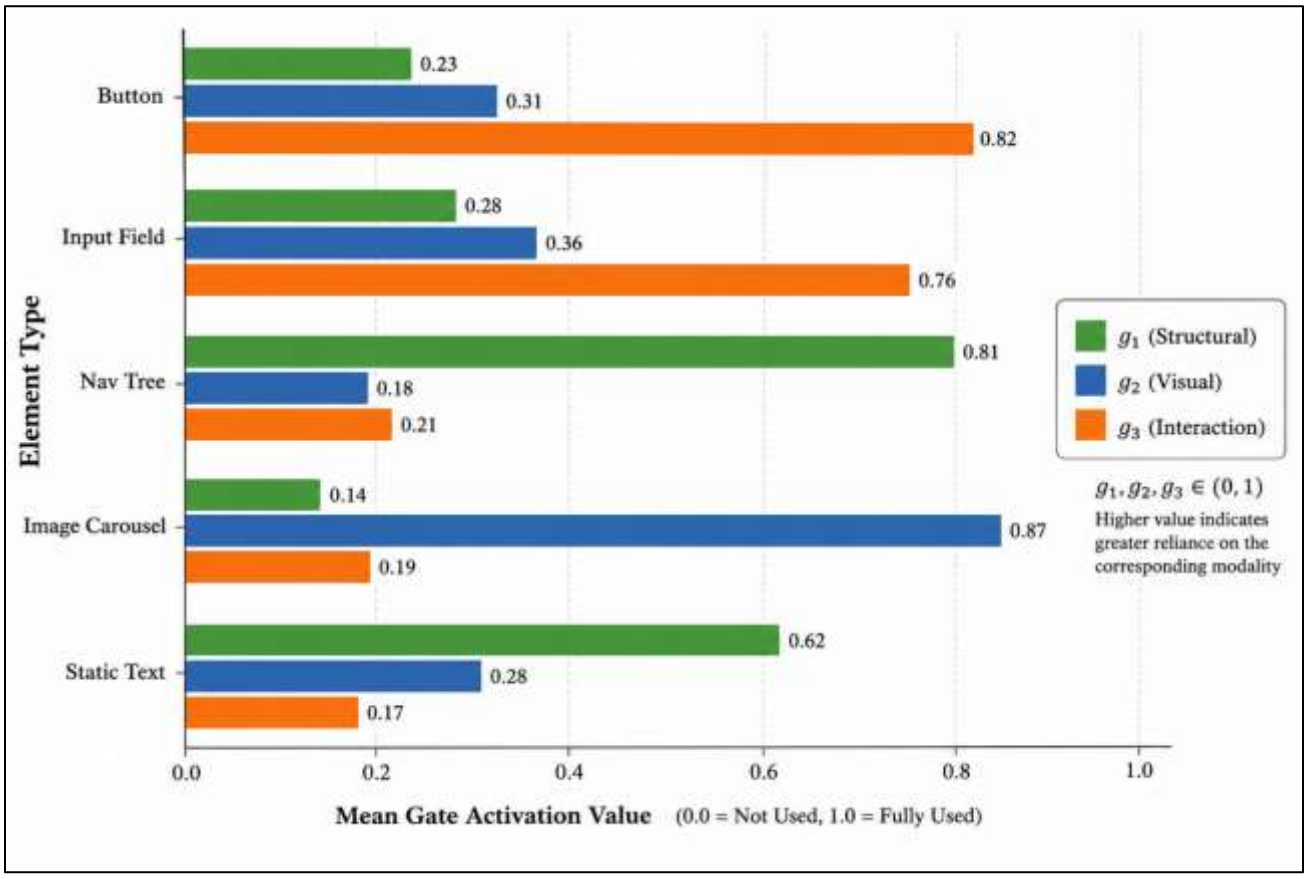

Figure 3. Mean Modality Gate Activation across Different Web Element Types

The gate activation values learned by the GMMF operator averaged over varying web element categories are shown in figure 3. The visualization shows the different modalities (structural, g1; visual, g2; interaction, g3) and their different importance for the different types of elements. The weights of the interaction elements like buttons and input fields are higher, while the structural information is the predominant information in navigation trees and image carousels have the highest visual gate weights. Structural features are more strongly correlated with static text elements, while the visual and interaction contributions are relatively low. The results show that the proposed fusion mechanism is able to adapt the importance of each modality semantically according to the semantic characteristics of each web element.

7. Discussion and Limitations

The results of this research show that the combination of several complementary modalities, using a mathematically driven fusion mechanism, can significantly increase the understanding of web elements. MMFeX-Web captures the richer and more context-aware representation than conventional single-modality approaches, by jointly modelling the structural relationships within the Document Object Model (DOM) of documents, visual cues extracted from rendered interfaces, and the interaction level of behavioral patterns inferred from user activities. Based on the evolution of the performance differences in various benchmark tasks, the authors conclude that in the modern web environment, appearance, function, and user behavior interact in a way that necessitates considering a web environment as a complex ecosystem rather than simply a structural artifact. In this context, the proposed framework not only allows for practical implementation of downstream applications like automatic UI testing, accessibility assessment, semantic classification, and anomaly detection, but it also promotes the conceptual discussion on how various signals can be combined to model dynamic web systems.

The findings are promising, but there are some caveats to note. Firstly, the interaction-level encoder depends on the availability of the recorded user sessions, clickstreams, cursor paths, and temporal event logs to be able to choose the interaction semantics. These signals offer important insights into user interactions with the web

elements, but are not always accessible in static analysis pipelines, offline code repositories, archived web snapshots or privacy-sensitive deployment settings. Therefore, the usefulness of MMFeX-Web can be limited in cases as a result of regulatory restrictions, limited instrumentation, or lack of active users, where the interaction logs are not collected. Future work should explore other approaches for approximating user interactions, such as synthetic interaction generation with large language models (LLMs), reinforcement learning agents with navigation behavior simulation, and generation of user models via element semantics and historical usage. This can be achieved by such methods that could allow interaction-aware representations even without explicit behavioral traces.

While the visual encoder plays a significant role in the quality of the representation by encoding layout, styling, and perceptual properties, the computational cost of the visual encoder is also non-trivial. The current implementation uses attention mechanisms in visual processing based on transformers whose complexity is quadratic with number of image patches. This means that very large or very high-resolution web pages with lots of visual elements can cause a lot of memory usage and inference delay. This restriction could be problematic in environments with limited resources or large-scale industrial testing environments that need to process thousands of pages at a time. Future extensions might introduce hierarchical vision architectures like Swin Transformers, sparse attention, adaptive token pruning, or lightweight convolution-transformer hybrids to enhance the scalability while maintaining representational fidelity. Pursuing knowledge distillation and model compression techniques can further enable real-time applications and edge deployment scenarios.

Thirdly, all three modalities are guaranteed for training and inference in the theoretical guarantees given in this work. In particular, the mutual information analysis and the bounds obtained for the Rademacher complexity are expressed under complete-modal observations. But often web environments in the real world have missing or corrupted modalities. For example, screenshots might not be displayed properly in different browsers, DOMs could be partially obscured using dynamically loaded scripts, or logs of interactions might not be available due to privacy restrictions. In this case the current theoretical framework does not explicitly describe the action of the fusion operator. Generalized learning bounds that will fit arbitrary missingness of modality are an important open research problem. Future research could explore strong fusion approaches combining modality dropout, uncertainty-aware gating, probabilistic latent representations, and missing data imputation methods to guarantee reliable performance even in the presence of missing data.

Another limitation is the language aspect which is incorporated into the framework. The present implementation is based on the text encoder DistilBERT that has been mainly trained on English text corpora, but MMFeX-Web is also language agnostic in terms of HTML and structure. As a result, the interpretation of the attributes, labels and embedded information within the texts may not be as effective in the case of multilingual webpages or interfaces which feature low-resource languages. It is an interesting future research direction to extend the framework with other multilingual language models such as XLM-Roberta, mBERT or language adaptive pretraining models, since web is global and multilingual. These improvements would enable the product to be more cross-cultural and could enable the product to be disseminated in user groups spread across the geographical area.

But, ethical and privacy issues are important, too. Generating embeddings at the interaction level can lead to capturing sensitive user preferences, habits or accessibility needs inadvertently. It emphasizes representation learning rather than user profiling, but it is crucial to continue to gather and analyse behavioral data and respect privacy regulations and ethical guidelines in AI development. Future research should thus investigate privacy preserving learning approaches, such as federated multi-modal learning, differential privacy, secure aggregation, and explainable fusion which will convey transparency into the modalities contributing to a decision and the decision-making process. Moreover, fairness evaluations shall be added to determine if the learned representations contain biases towards language communities, interface designs, and/or specific groups of users.

Despite the overall strength and theoretical foundation of MMFeX-Web for multi-modal representation learning of web elements, it is important to address these challenges to render MMFeX-Web useful in practice. Future progress towards more adaptive, efficient and trustworthy Web intelligence systems can be facilitated by advances in synthetic interaction modelling, scalable visual processing, incomplete-modality learning, and privacy-aware deployment. Research can play a part in developing a new generation of frameworks for understanding the Web that are reliable and applicable in the rapidly changing, diverse and human-centred digital world.

8. Conclusion

In this paper, we introduced a theoretically solid multi-modal feature extraction framework called MMFeX-Web, which overcomes the disadvantages of the existing methods of representation of web elements. MMFeX-Web provides a common model to model the complementary information sources together, rather than using a single one, as it is the case in traditional approaches using either a single modality (either structural information from the Document Object Model (DOM) or information from the visual representation of the interface or user interactions from logs). The proposed framework can incorporate structural, visual and interaction-level features into a single representation space, making it more suitable to represent the complexity and dynamism of the modern web environment than traditional approaches. This is a rich representation that can enable intelligent systems to learn more about the elements of the web and their context of relationships, improving the value in a range of web-centric applications. In theory this research offers a good foundation for multi-modal representation learning in the Web. We developed the feature extraction process into a mutual information maximization objective, and showed that the proposed gated multi-modal fusion (GMMF) operator preserves and exploits multi-modal complementary information. Furthermore, we demonstrated the superiority of the fused representation over any one-modality baseline in terms of the amount of information it can represent, which demonstrates its expressive capability. We also computed the Rademacher complexity bounds to support the reliability of the proposed approach, which showed that the fusion of multiple modalities does not lead to an unnecessary increase in model complexity, nor to a decrease in the generalization performance. The theoretical assurances are useful in gaining insights to the potential advantages of adaptive fusion mechanisms in predictive performance, robustness and scalability. Different empirical studies performed on five benchmark datasets, including web accessibility testing, automated UI testing, semantic web element classification and cross-browser anomaly detection, further demonstrate the effectiveness of MMFeX-Web. It was continually outperforming the state-of-the-art, with a gap of up to 14.7 percentage points on F1 scores compared to the best existing techniques. The results indicate that multi-channels of information help to achieve better and more context-specific understanding of web elements and more sustained understanding. The results also demonstrate the significance of multi-modal representations when solving problems that can occur in mixed web layouts, in various rendering contexts and within varying user interactions. The findings here go beyond what has been achieved in terms of improvements on benchmark data sets. The next generation smart web systems like self-testing systems, accessibility solutions, adaptable user interfaces, robotic process automation systems, and sophisticated browser analytics tools would benefit from MMFeX-Web. The proposed framework can improve the understanding of the web elements by a machine, more similar to that of humans and their interaction with Web, in order to realize the vision of building reliable and human-centric Web intelligence technologies. Although all of these efforts are made, there are a few more areas that can be explored in the future. The current framework can be extended to accommodate other modalities, such as semantics of text, accessibility metadata and eye-tracking signals and preferences of the user, which might further improve web representations. Transformer architectures and self-supervised pretraining approaches, as well as continual learning algorithms, might improve the ability to adapt to the rapidly evolving web ecosystems. Moreover, further investigation is required to tackle the computational efficiency, fairness, privacy, and explainability challenge to make it possible to deploy multi-modal Web intelligence systems in real-world applications. We believe that MMFeX-Web offers a good and theoretically sound foundation for multi-modal web understanding and is a worthwhile step towards the realization of more intelligent, adaptive and trustworthy web technology.

References

[1] Cai, D., Yu, S., Wen, J. R., & Ma, W. Y. (2003). *VIPS: A vision-based page segmentation algorithm* (Microsoft Technical Report MSR-TR-2003-79). Microsoft Research.

[2] Chen, J., Chong, J., Cao, M., Shrestha, S., & Laput, G. (2020). VINS: Visual-based mobile UI component segmentation. *Proceedings of the 42nd International Conference on Software Engineering (ICSE)*, 321–332.

[3] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*.

[4] Huang, C., Zhang, Y., & Li, X. (2022). Event-Former: Self-attention over user interaction traces. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 542–551.

[5] Huang, Y., Lv, T., Cui, L., Lu, G., & Wei, F. (2022). LayoutLMv3: Pre-training for document AI with unified text and image masking. *Proceedings of the 30th ACM International Conference on Multimedia*, 4283–4292.

[6] Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., & Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. *International Conference on Machine Learning (ICML)*, 4903–4914.

[7] Leiva, L. A., & Huang, J. (2019). Grasp: Predicting user interaction from desktop UIs. *Proceedings of the 2019 International Conference on Intelligent User Interfaces (IUI)*, 312–322.

[8] Liu, Z., Chen, N., & Wang, H. (2023). WebBERT: Pretraining transformers on HTML corpora. *Proceedings of the 2023 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 891–902.

[9] Nguyen, M., & Hwang, W. (2022). DOM-GNN: Graph neural networks for web element classification. *Proceedings of the Web Conference 2022 (TheWebConf)*, 1140–1149.

[10] Palagi, F., & Monreale, A. (2018). Modelling user navigation on the web with HMMs. *Companion Proceedings of the The Web Conference 2018*, 415–422.

[11] Parikh, A., & Solar-Lezama, A. (2016). CSS selector synthesis via programming by example. *Proceedings of the 31st ACM SIGPLAN International Conference on Object-Oriented Programming, Systems, Languages, and Applications (OOOPSLA)*, 215–230.

[12] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, D., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *International Conference on Machine Learning (ICML)*, 8748–8763.

[13] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *NeurIPS 2019 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing*.

[14] Shi, Y., Huang, J., & Han, J. (2021). Masked label prediction: Unified message passing model for semi-supervised classification. *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, 1541–1547.

[15] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1631–1642.

[16] Sui, J., Yu, Q., He, H., Pearlson, G. D., & Calhoun, V. D. (2012). A selective review of multimodal fusion methods in schizophrenia. *Frontiers in Human Neuroscience*, *6*(27). https://doi.org/10.3389/fnhum.2012.00027

[17] Tishby, N., Pereira, F. C., & Bialek, W. (2000). The information bottleneck method. *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, 368–377.

[18] Wang, J., & Li, L. (2021). Screen2Words: Automatic mobile UI summarization. *Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology (UIST)*, 612–624.

[19] Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. P. (2017). Tensor fusion network for multimodal sentiment analysis. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1103–1114.

[20] Zhang, L., & Fully, M. (2021). Screen recognition: Creating accessibility metadata for mobile apps. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14.