



Modeling and understanding user interfaces from pixels

Amanda Sweeny

8th Summer School on Computational Interaction | Apple | June 4th, 2024

About Me

Amanda Swearngin



Apple, AI/ML

2019+: Research Scientist at Apple

2015-2019: Ph.D. from University of Washington

“Expanding Interface Design Capabilities through Semantic and Data-Driven Analyses”

- ML + Software Engineering to improve tools for UX and UI design

2012 - 2015: Software Engineer, Microsoft

Undergrad + Masters: University of Nebraska-Lincoln

About Me

Amanda Swearngin

My Work at Apple

- UI Understanding
- Accessibility
- Developer Tools

Agenda

Modeling UIs from Pixels (UI Understanding)

Examples

Practical Exercise: Training a screen recognizer and other UI understanding models (screen classification, screen similarity)

Next Generation of UI Understanding (LLMs+VLMs)

Examples

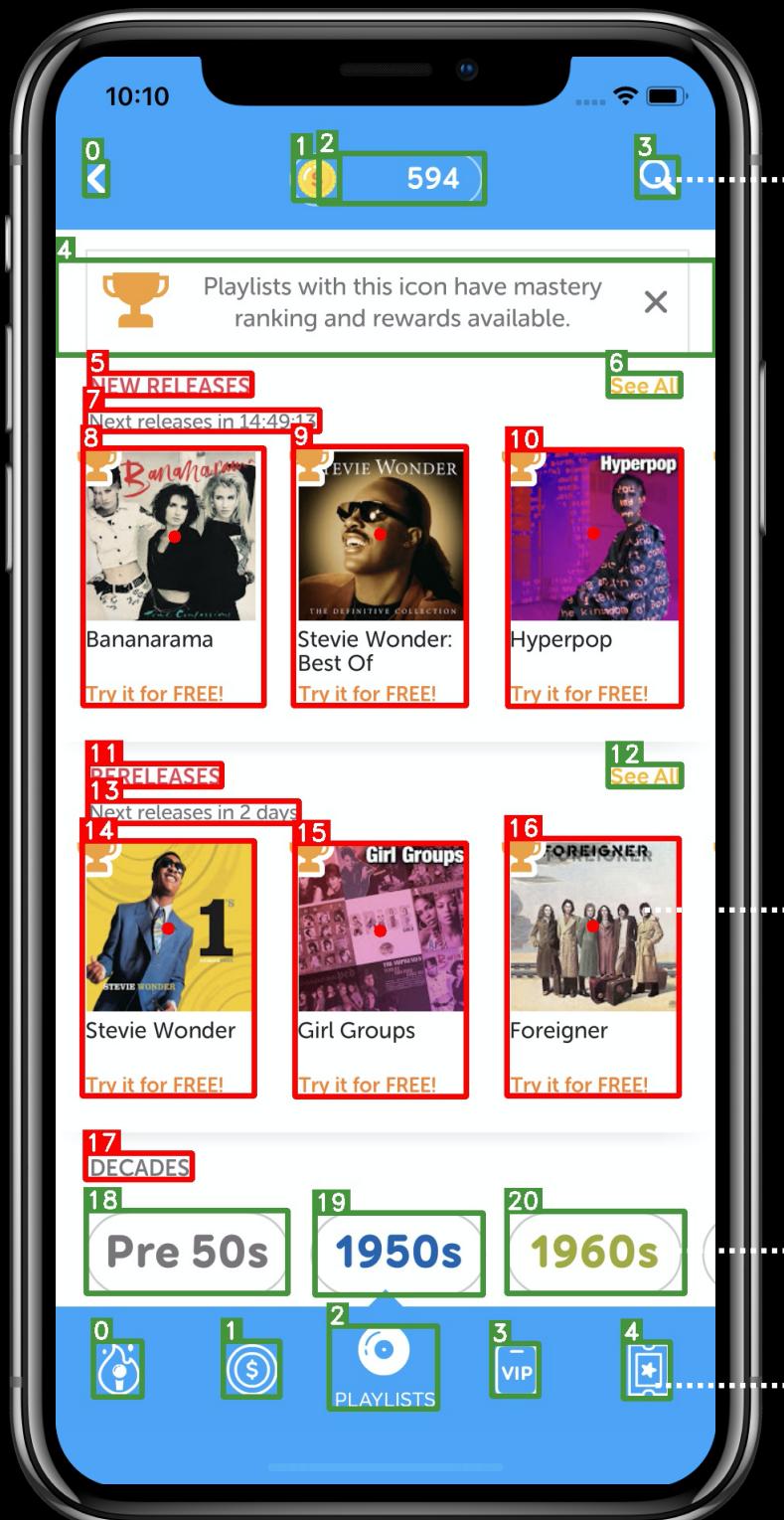
Practical Exercise: Use LLM & VLM for screen Q&A

UI Understanding

Goal

Give machines human-level abilities
to understand and interact with
user interfaces

UI Understanding



Icon: Search

Group: Picture + Title + Text

Text: Clickable

Tab Bar Icon: Fire

Applications

Accessibility

Automation

UI Testing

UI Design

UI Agents

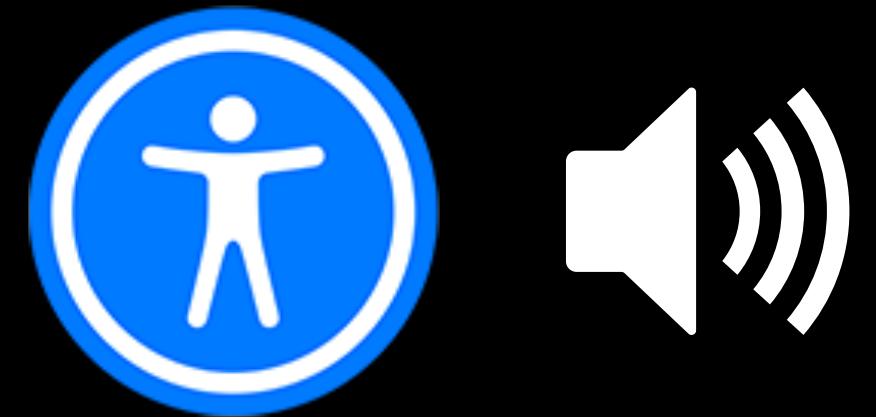
...

Examples

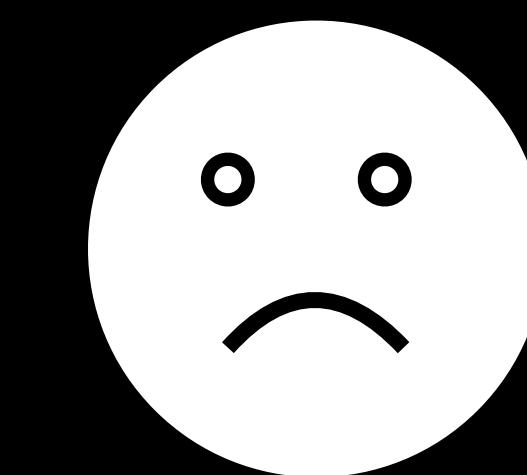
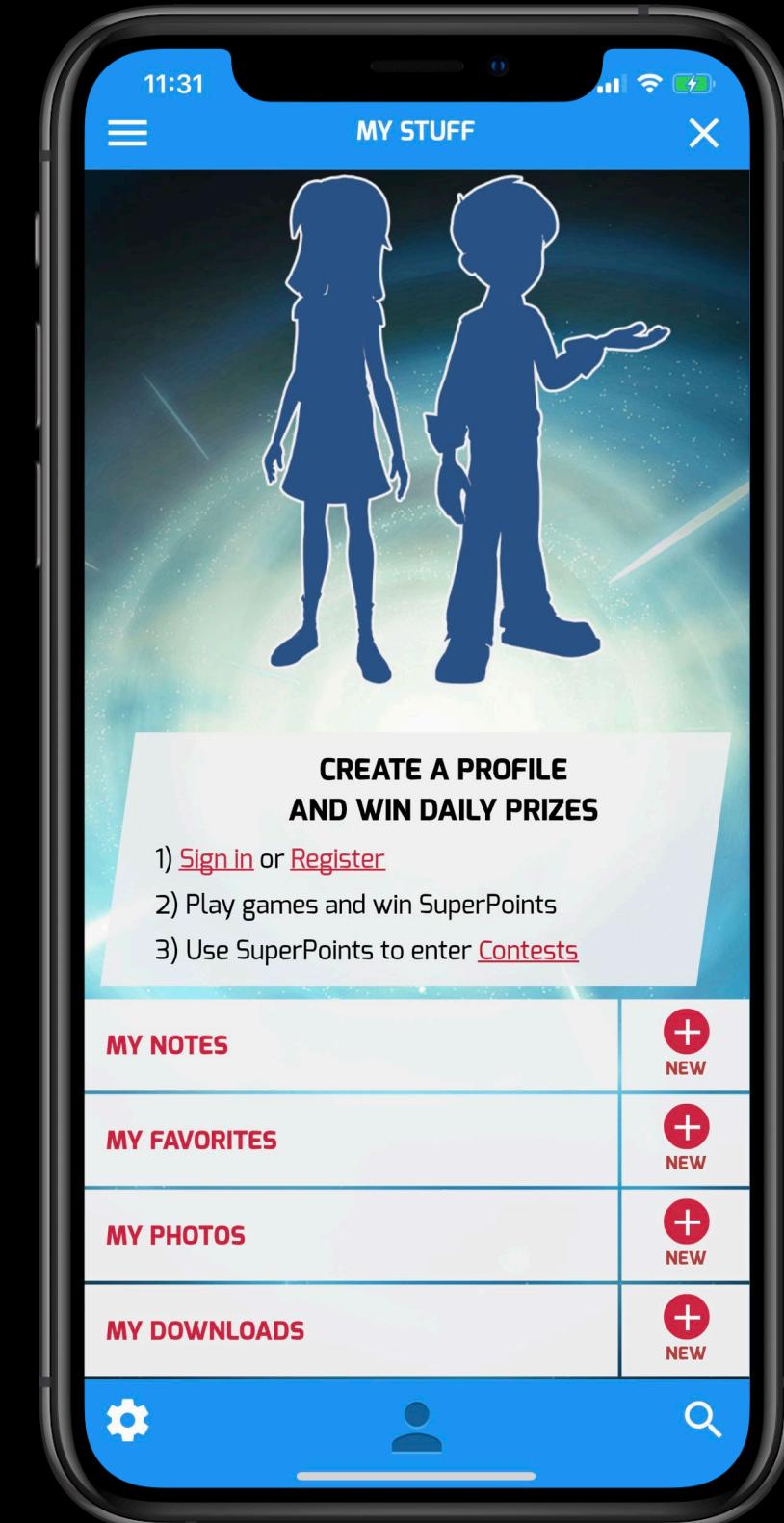
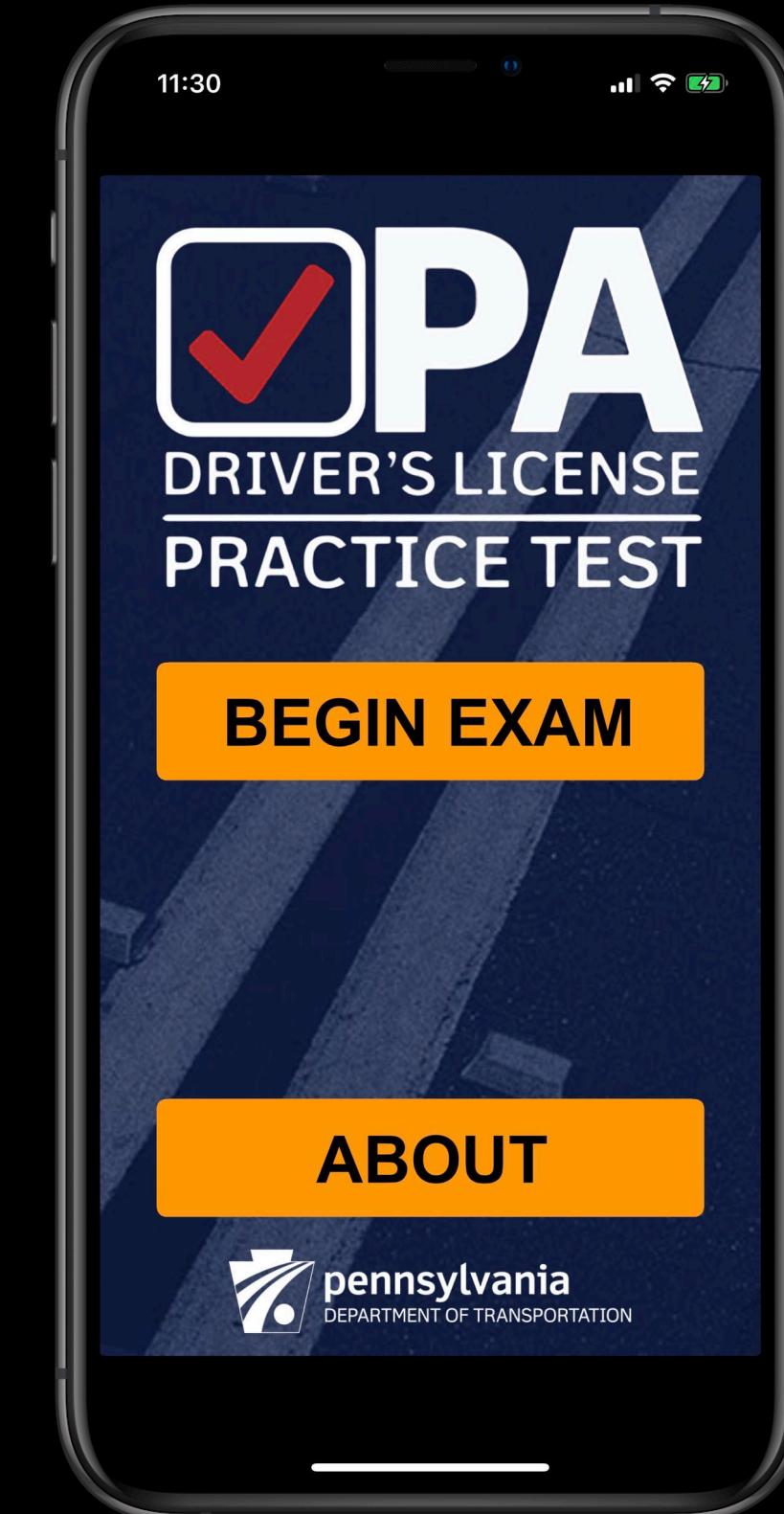
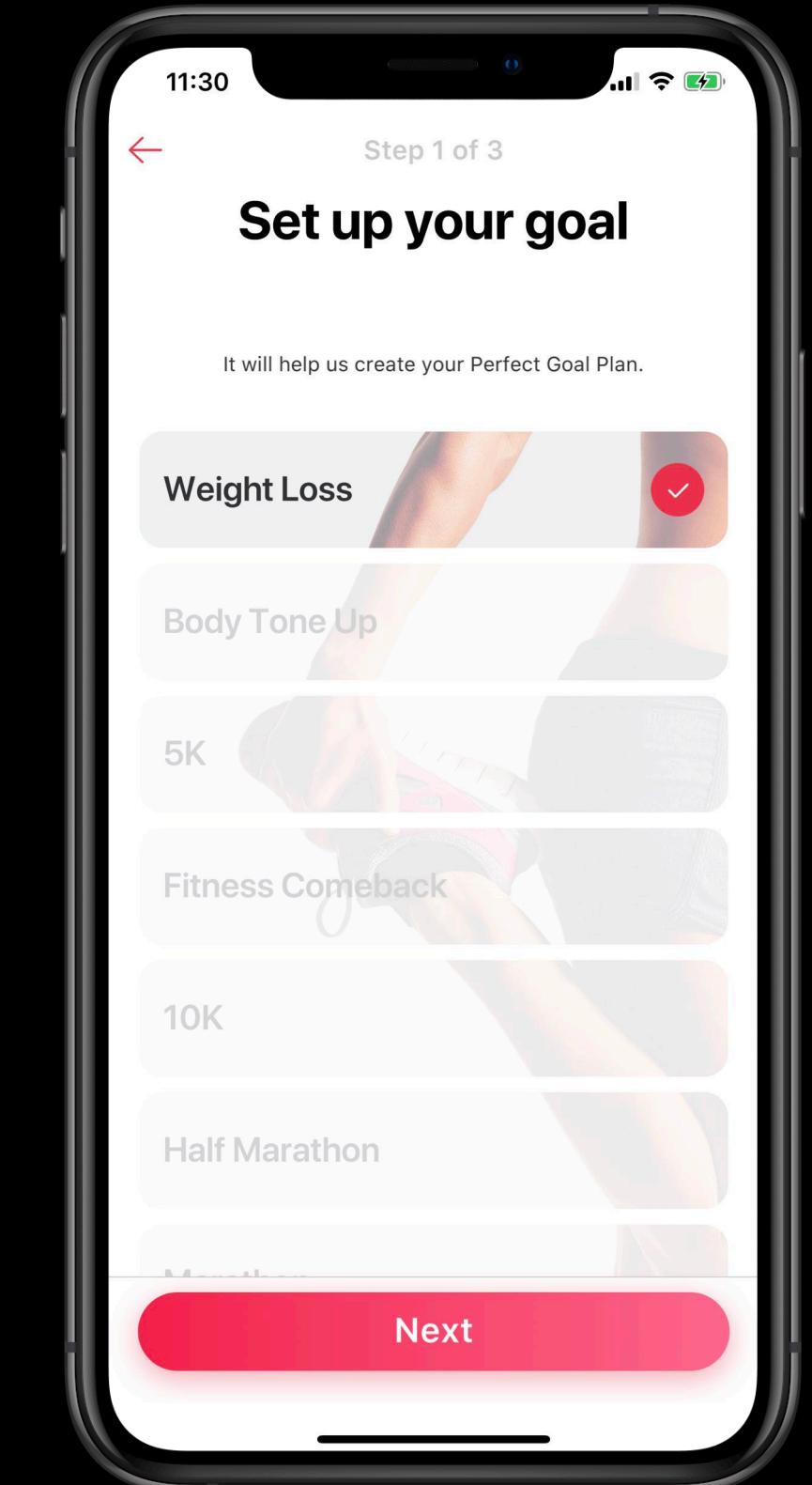
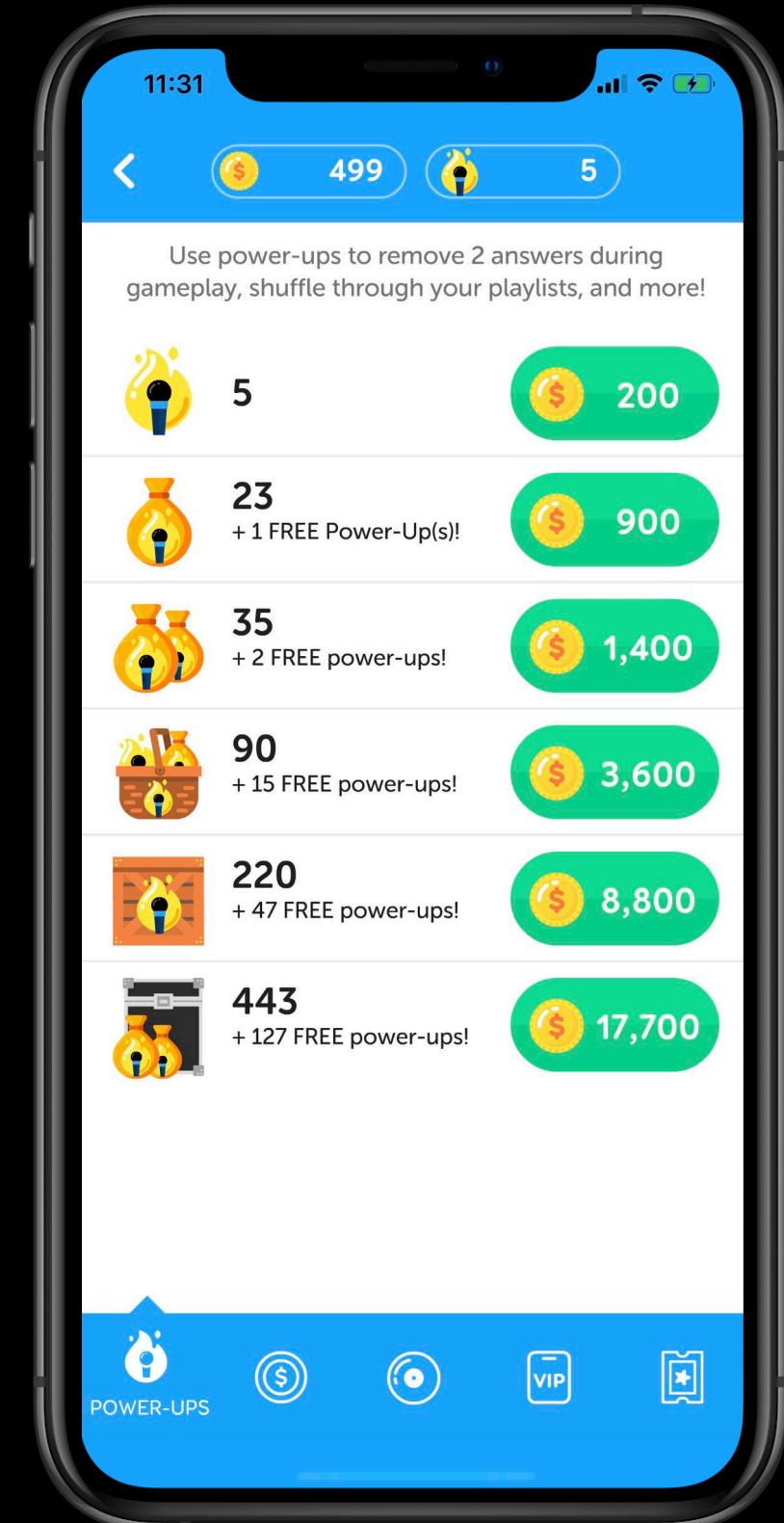
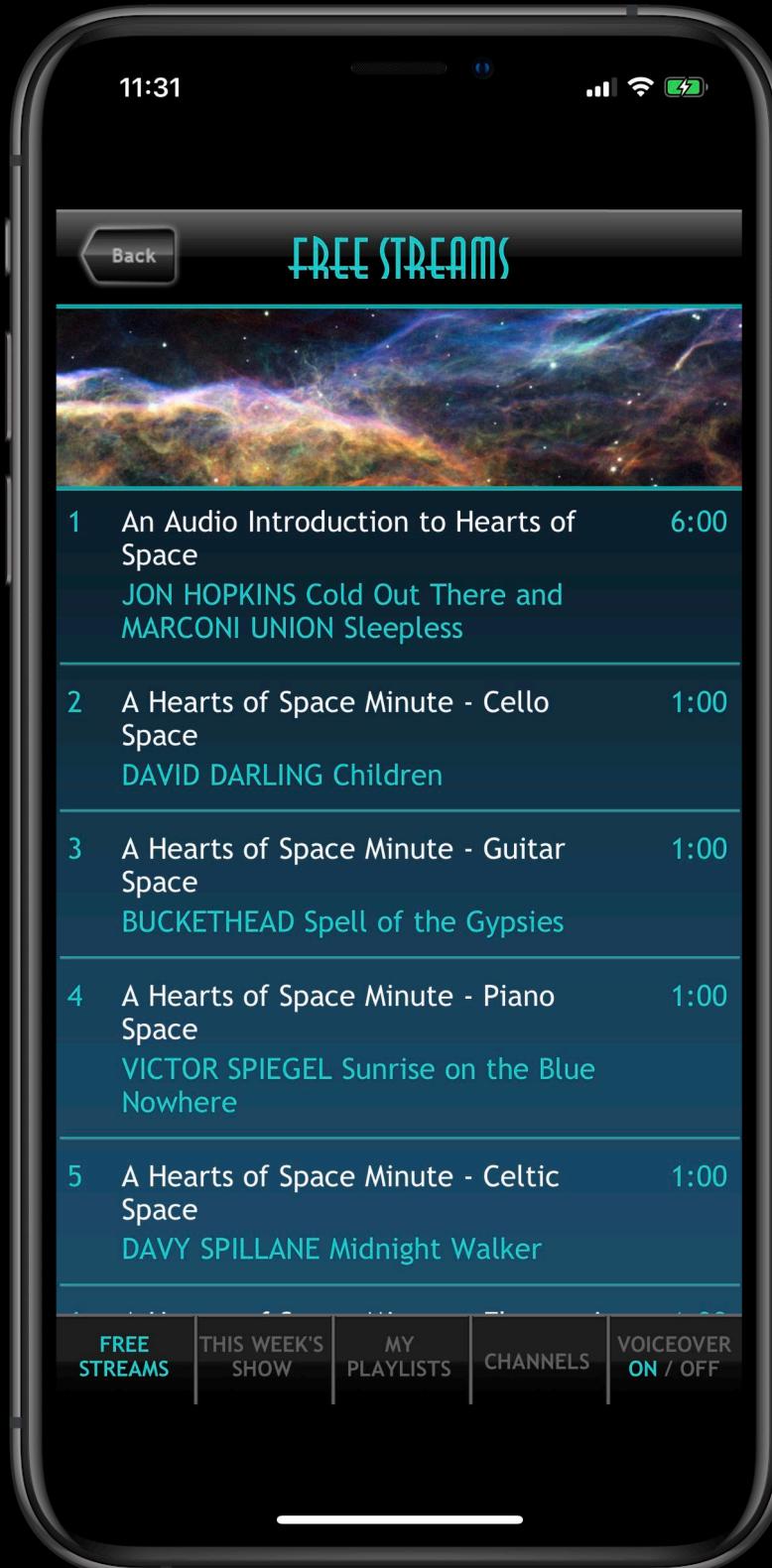
Screen Recognition

Creating Accessibility Metadata for Mobile Apps from Pixels¹

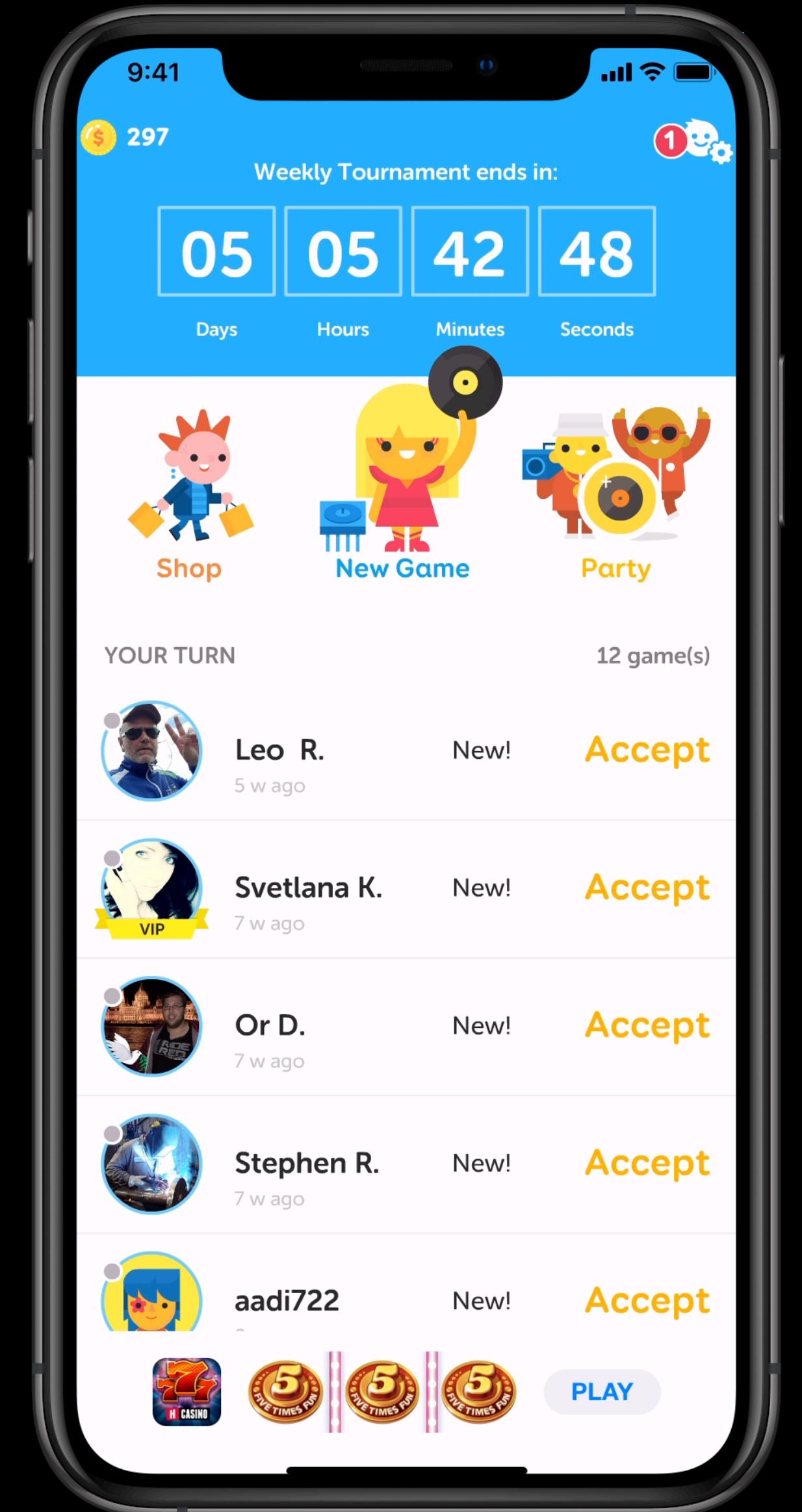
1. Xiaoyi Zhang, Lilian De Greef, Amanda Swearngin, Samuel White, Kyle Murray, Lisa Yu, Qi Shan, Jeffrey Nichols, Jason Wu, Chris Fleizach, Aaron Everitt, Jeffrey P Bigham, "Screen recognition: Creating accessibility metadata for mobile applications from pixels." Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 2021.



```
hierarchy": {  
  "accessibilityElements": [ ...  
  "class": "Text",  
  "frame": [10,10,150,210]  
  "label": "Accessibility News"  
}  
...
```

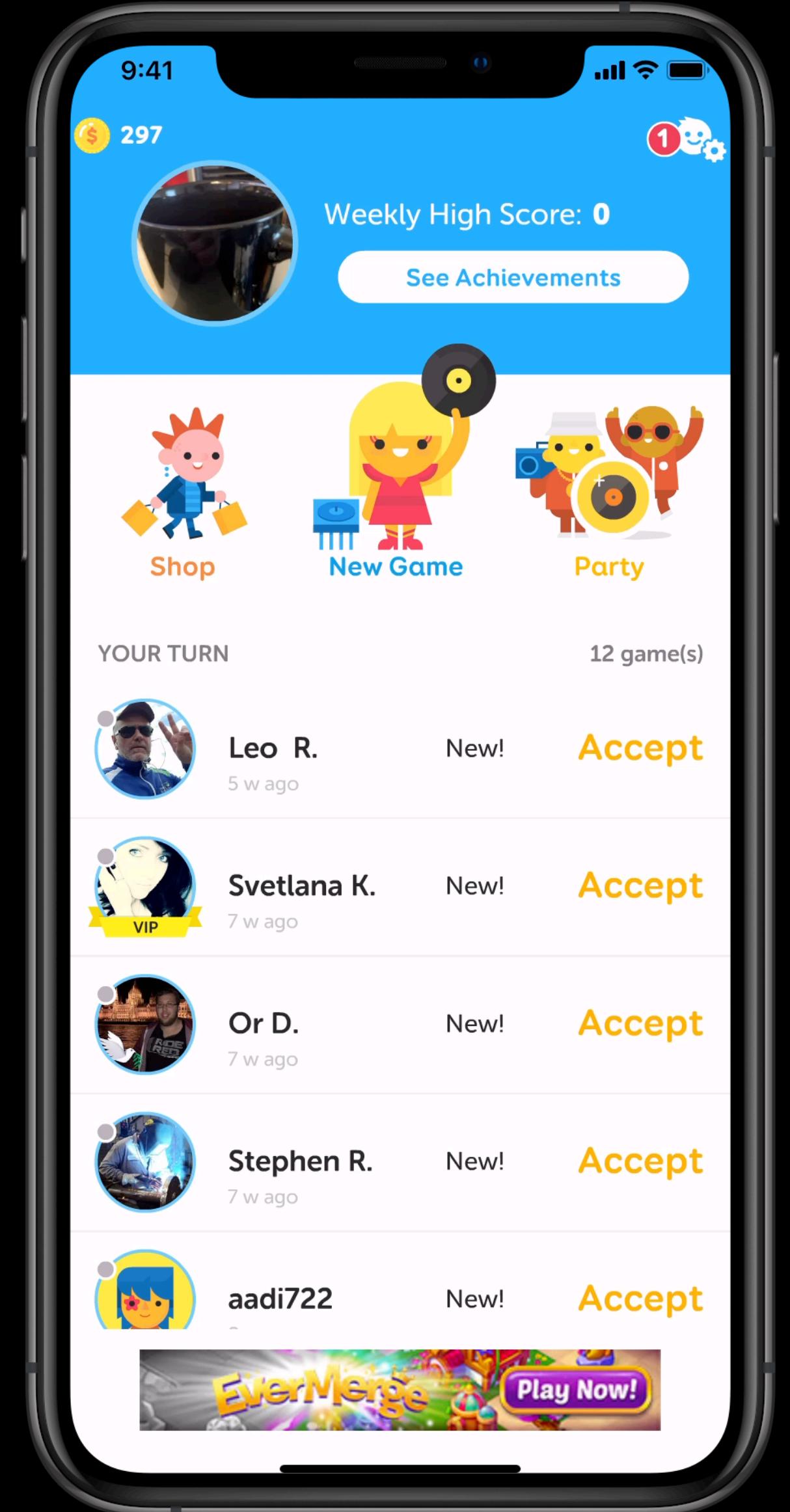


Using Regular VoiceOver



This empty feedback sound means there is no accessibility element.

Using Screen Recognition Enabled VoiceOver



Colored borders are
for demonstration
purposes only.

“I am so excited I just NEED to share. Last night I downloaded SongPop2... which has no AX elements and is completely unusable. So I decided to enable [Screen Recognition] that Aaron demoed yesterday. Guess who has a new high score???!!! I am in AWE! This is incredible. It is the first mainstream game I've ever been able to play on iOS besides Trivia Crack.”

— Accessibility QA Engineer

Screen Recognition: Creating Accessibility Metadata for Mobile Applications from Pixels

| | | |
|--|--|--|
| Xiaoyi Zhang Apple Inc. xiaoyiz@apple.com | Lilian de Greef Apple Inc. ldegreef@apple.com | Amanda Sweenin Apple Inc. asweenin@apple.com |
| Samuel White Apple Inc. samuel_white@apple.com | Kyle Murray Apple Inc. kyle_murray@apple.com | Lisa Yu Apple Inc. lixiu_yu@apple.com |
| Qi Shan Apple Inc. qshan@apple.com | Jeffrey Nichols Apple Inc. jwnichols@apple.com | Jason Wu Apple Inc. jason_wu2@apple.com |
| Chris Fleizach Apple Inc. cfleizach@apple.com | Aaron Everitt Apple Inc. aeveritt@apple.com | Jeffrey P. Bigham Apple Inc. jbigham@apple.com |

ABSTRACT
 Many accessibility features available on mobile platforms require applications (apps) to provide complete and accurate metadata describing user interface (UI) components. Unfortunately, many apps do not provide sufficient metadata for accessibility features to work as expected. In this paper, we explore inferring accessibility metadata for mobile apps from their pixels, as the visual interfaces often best reflect an app's full functionality. We trained a robust, fast, memory-efficient, on-device model to detect UI elements using a dataset of 77,637 screens (from 4,068 iPhone apps) that we collected and annotated. To further improve detections and add semantic information, we introduced heuristics (e.g., UI grouping and ordering) and additional models (e.g., recognize UI content, state, interactivity). We built Screen Recognition to generate accessibility metadata to augment iOS VoiceOver. In a study with 9 screen reader users, we validated that our approach improves the accessibility of existing mobile apps, enabling even previously inaccessible apps to be used.

CCS CONCEPTS
 • Human-centered computing → Accessibility technologies.

KEYWORDS
 mobile accessibility, accessibility enhancement, ui detection

ACM Reference Format:
 Xiaoyi Zhang, Lilian de Greef, Amanda Sweenin, Samuel White, Kyle Murray, Lisa Yu, Qi Shan, Jeffrey Nichols, Jason Wu, Chris Fleizach, Aaron Everitt, Jeffrey P. Bigham. 2021. Screen Recognition: Creating Accessibility Metadata for Mobile Applications from Pixels. In *CHI Conference on Human Factors in Computing Systems (CHI '21)*, May 8–13, 2021, Yokohama, Japan. ACM, New York, NY, USA, 15 pages. <https://doi.org/10.1145/3411764.3445186>

🏆 CHI Best Paper (Review 4.4 / 5)

iPhones can now automatically recognize and label buttons and UI features for blind users

Devin Coldewey @techcrunch / 8:30 AM PST • December 3, 2020

Comment



Voiceover recognition is the way of the future!

IOS 14.1 found great use for voiceover recognition!

BFGC_Rattlehead @OBSKHRattlehead · Mar 15
 Is it just me or HBO max is not accessible with voiceover on iOS?
 3 1 1

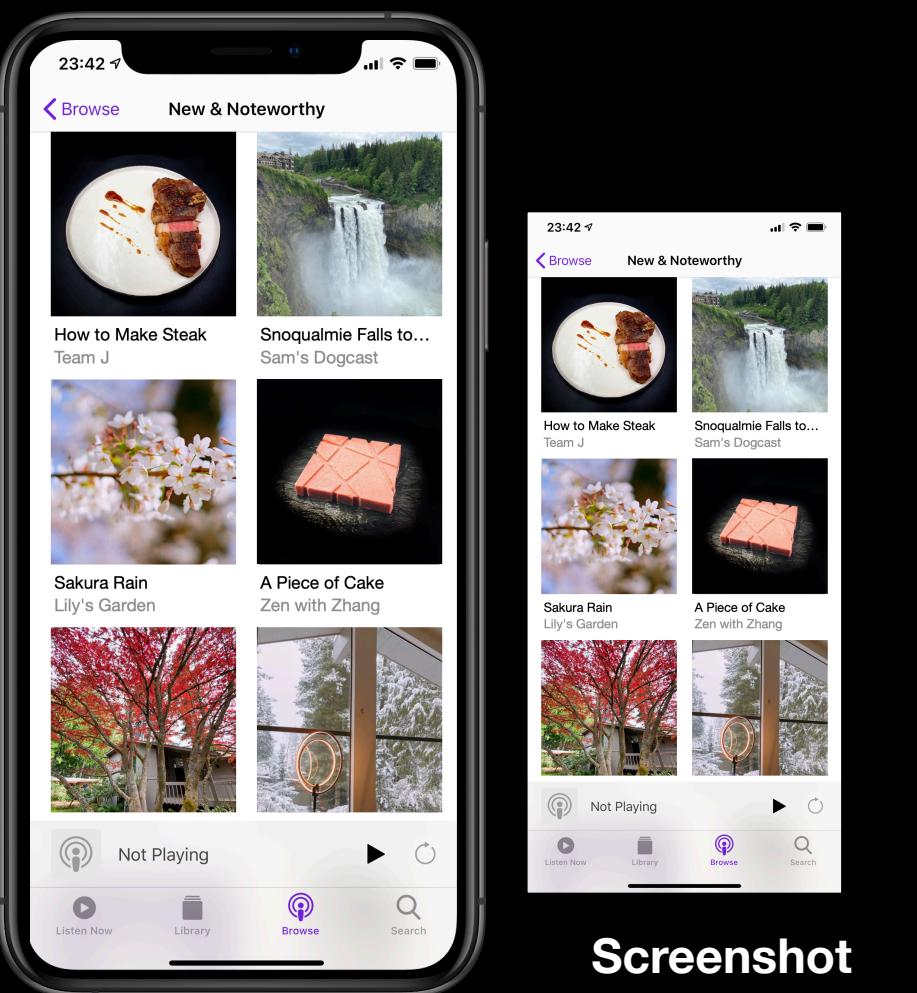
Allen Sale 🚶 @AstralAudio · Mar 15
 Try turning screen recognition on.

顾伶磊 V
 2020-9-21 16:33 来自 iPhone客户端
 视频当中听到的灯灯灯的声音就是我在使用旁白触摸屏幕，由于旁白抓取不到任何页面内容，从而返回了错误提示音。可以看到，当我们开启旁白识别后，旁白不仅正常读出了页面内容，还利用机器学习技术，正确的识别出了部分控件的控件类型。使得视障者可以更好地对页面进行操作。

Media Coverage

User Feedback

1) Data Collection



200,000+ Screens & UI Trees
8000+ iPhone & iPad apps

```
"hierarchy": {  
  "accessibilityElements": [ ...  
    "class": "UIButton",  
    "frame": [10,10, 150, 210]  
    ...  
  ]  
  ...  
}
```

UI Tree

Your Shopping Cart is empty.

Your Shopping Cart lives to serve. Give it purpose — fill it with books, CDs, DVDs, toys, electronics, and more. If you already have an account, [Sign In](#) to see your Cart.

Continue shopping on the [Amazon.com homepage](#), learn about [today's deals](#), or visit your [Wish List](#).

Sign in for the best experience

Returns are easy

Deal of the Day

OfferUp-com.offerup.iphone.consumer-1551985444712.json Modified

```
      "height" : 896
    }
  },
  "accessibilityNavigationStyle" : "UIAccessibilityNavigationStyle$",
  "AMPSuperclassChain" : [
    "UIView",
    "UIResponder",
    "NSObject"
  ],
  "AMPPrivateAccessibilityTraits" : [
    "accessibilityTraits" : [
      "UIAccessibilityTraitNone"
    ],
    "accessibilityViewIsModal" : false,
    "AMPObjec...uid" : 2,
    "accessibilityLabel" : "Edit profile",
    "shouldGroupAccessibilityChildren" : false,
    "isHidden" : true,
    "accessibilityIdentifier" : null,
    "kAXIsVisibleAttribute" : false,
    "accessibilityFrame" : {
      "origin" : {
        "x" : 100,
        "y" : 205
      },
      "size" : {
        "width" : 214,
        "height" : 396
      }
    }
  ]
}
```

Jane Smith

Celebrating Your Friendversary With Your Sister

Jane, you may have known each other longer, but 6 years ago today, you became friends with your sister, Jill, on Facebook! Here's a look at some memories you've shared together over the years.

+ Add Bio

Add Story Edit Profile Activity More

Messenger

Restaurants

Google

University of Washington

Parking Area E18

Husky Union Bldg (HUB)

University of Washington Seattle, WA 98195

Apt/Suite Optional

Instructions Optional

4:43 PM

No SIM WiFi 5:07 PM

Search Watch

Restaurants

No SIM WiFi 5:09 PM

Watchlist News Food Spo Sort by Open now Cuisine

CNN International nightline GREY'S ANATOMY W

Random Experiments Yesterday at 3:00 PM Make a "Matchless" Survival Fire Kit

TKOR Let's try using a bit of backyard science

1.6K 51 Comments 610 Shares Like Comment Share

Portage Bay Cafe - Roosevelt Airy spot for organic breakfast & lunch

THE BOY WHO HARNESSED THE WIND

BUDAPEST BREAKER UPPIERS

Hue white lamp 4 Unavailable Philips hue - 642945 Unavailable

Default Room

Accessories

Done

Q Friends

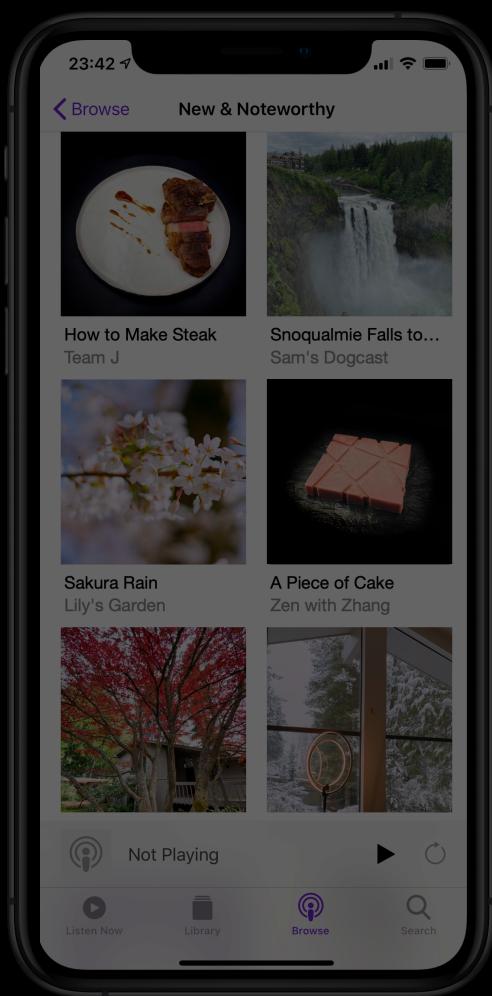
Johnmarc91 Tap to chat

Besties Opened - 13w

Apple Talktester1 ScreenShot - 19w

Jane Smith

1) Data Collection

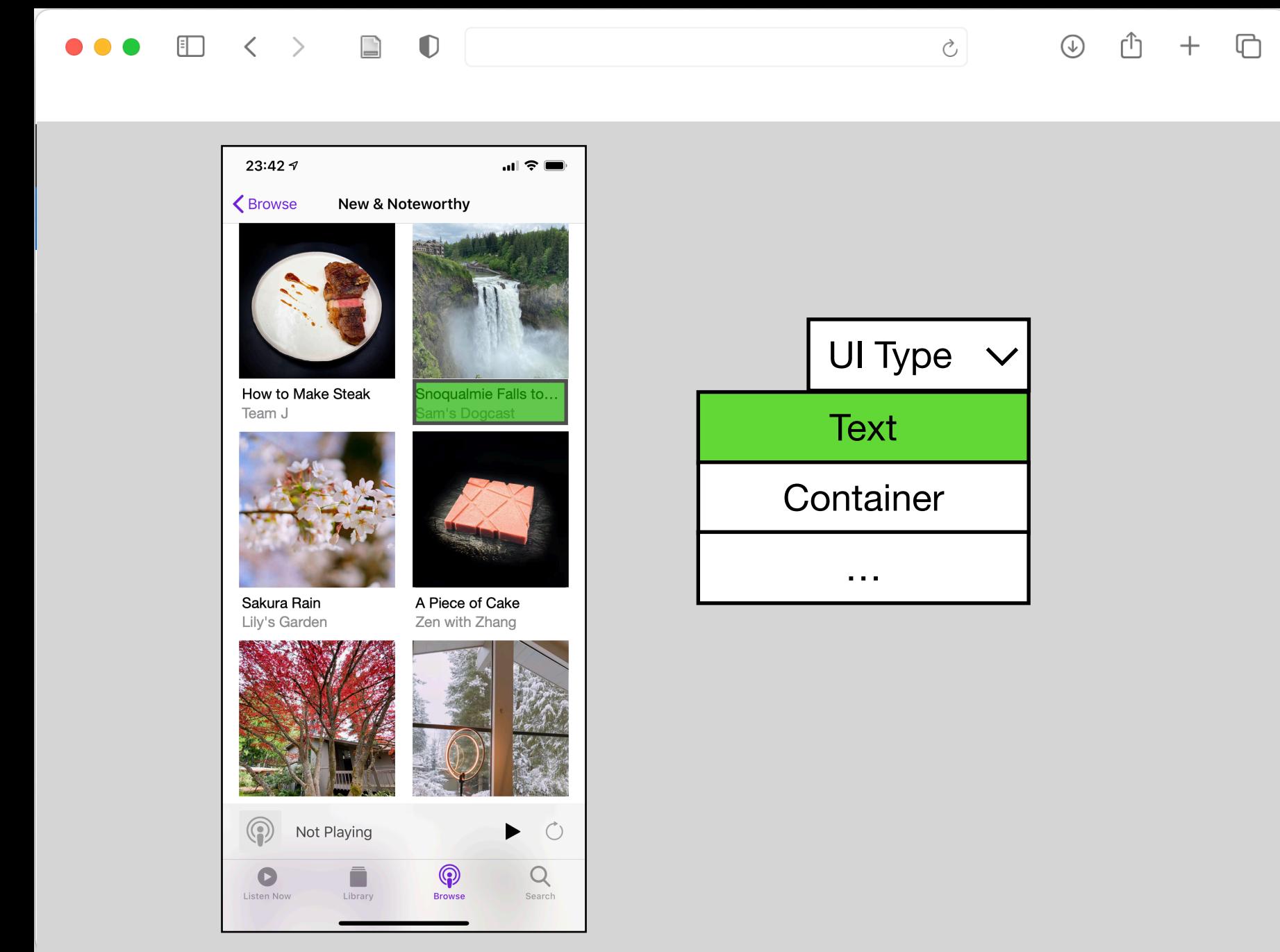
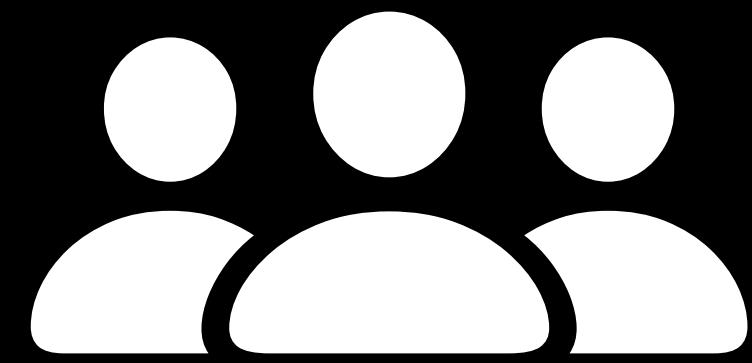


Screenshot

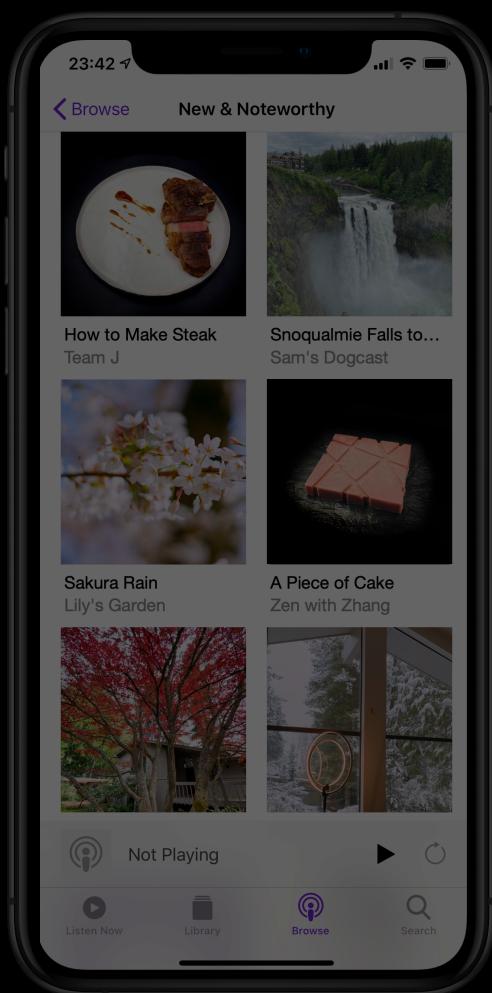
```
"hierarchy": {  
  "accessibilityElements": [ ...  
    {"class": "UIButton",  
     "frame": [10,10, 150, 210]  
    }  
  ]  
  ...  
}
```

UI Tree

2) Data Annotation

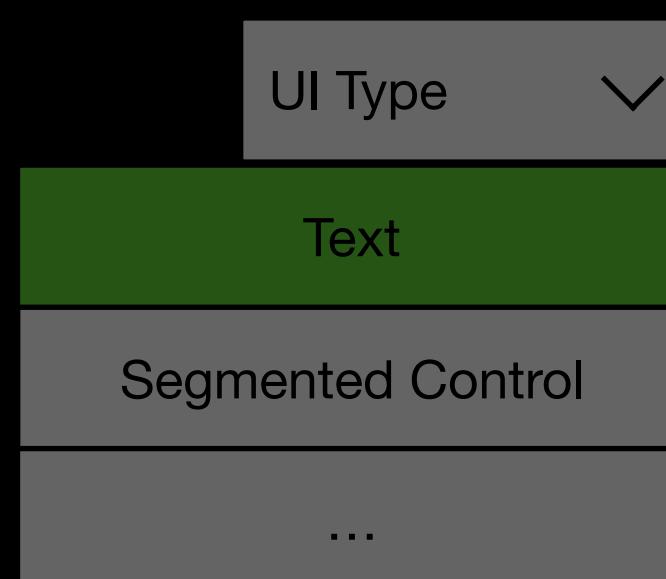
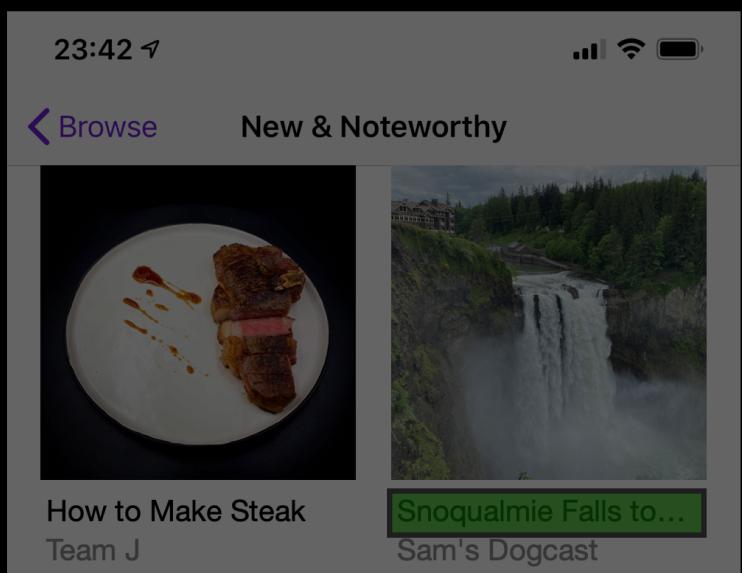


1) Data Collection

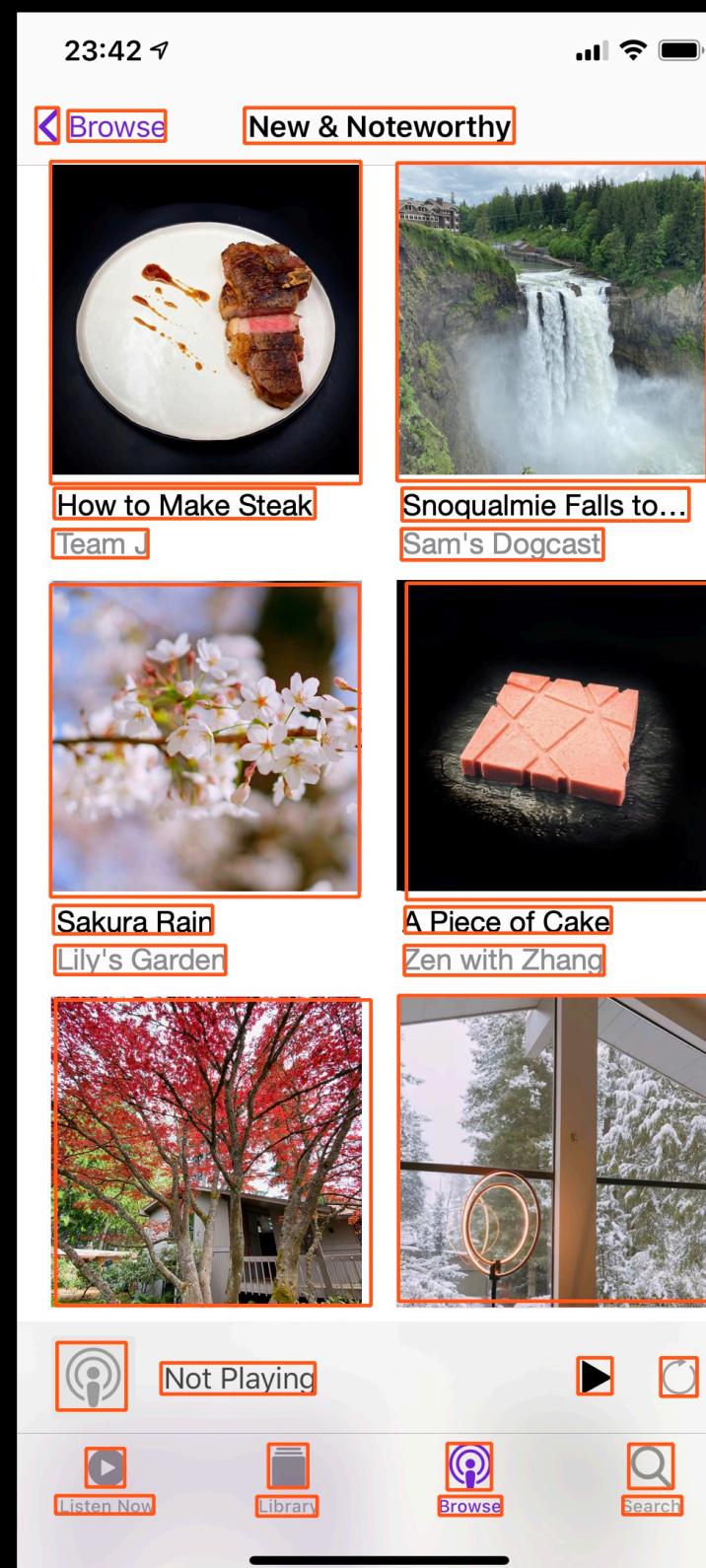


```
hierarchy": {  
  "accessibilityElements": [ ...  
    {"class": "UIButton",  
     "frame": [10,10, 150, 210]  
    }  
  ]  
  ...  
}  
  
UI Tree
```

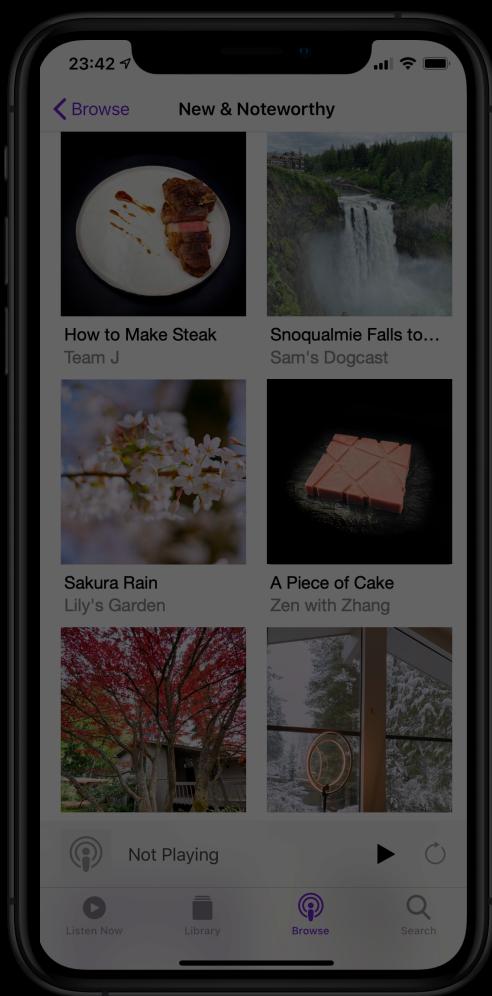
2) Data Annotation



3) UI Detection Model

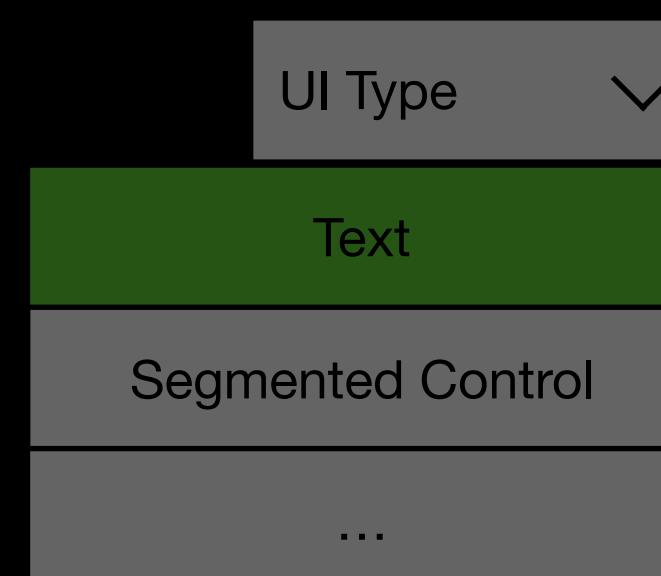
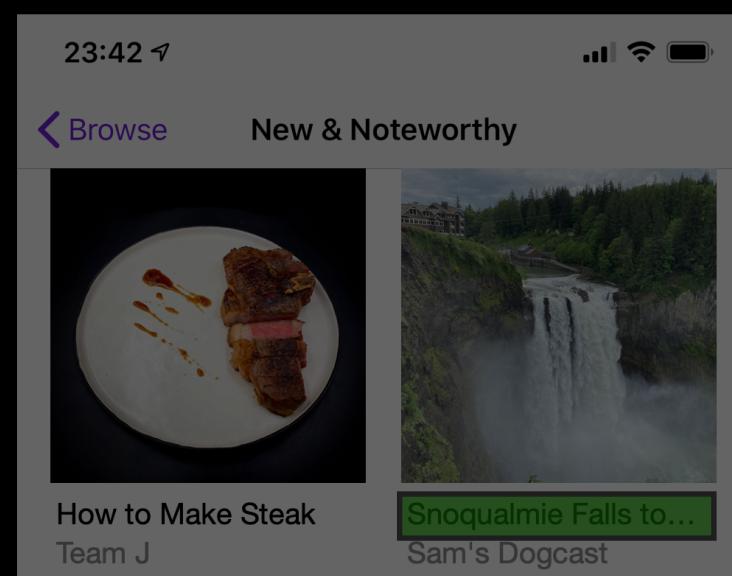


1) Data Collection

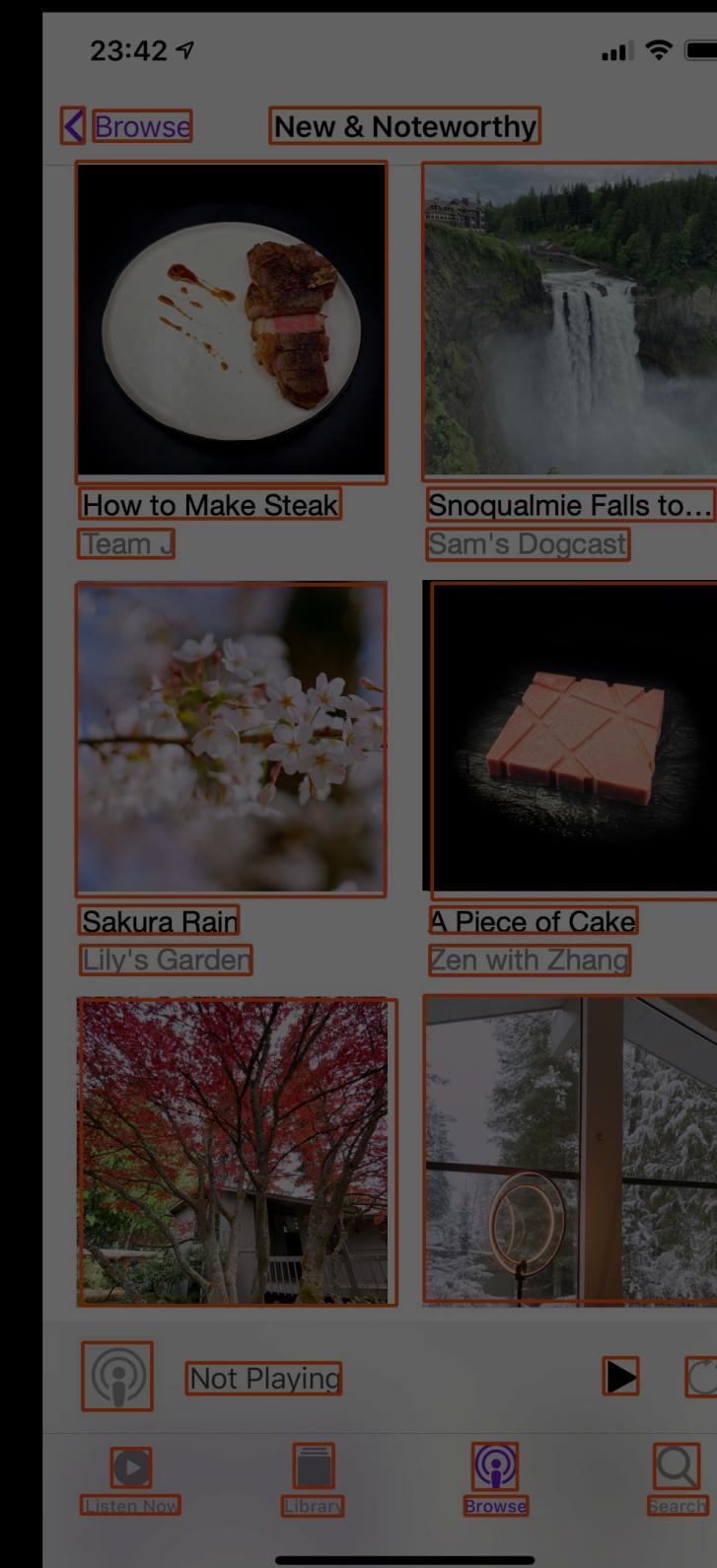


```
"hierarchy": {  
  "accessibilityElements": [ ...  
    {"class": "UIButton",  
     "frame": [10,10, 150, 210]  
    }  
  ]  
  ...  
}  
  
UI Tree
```

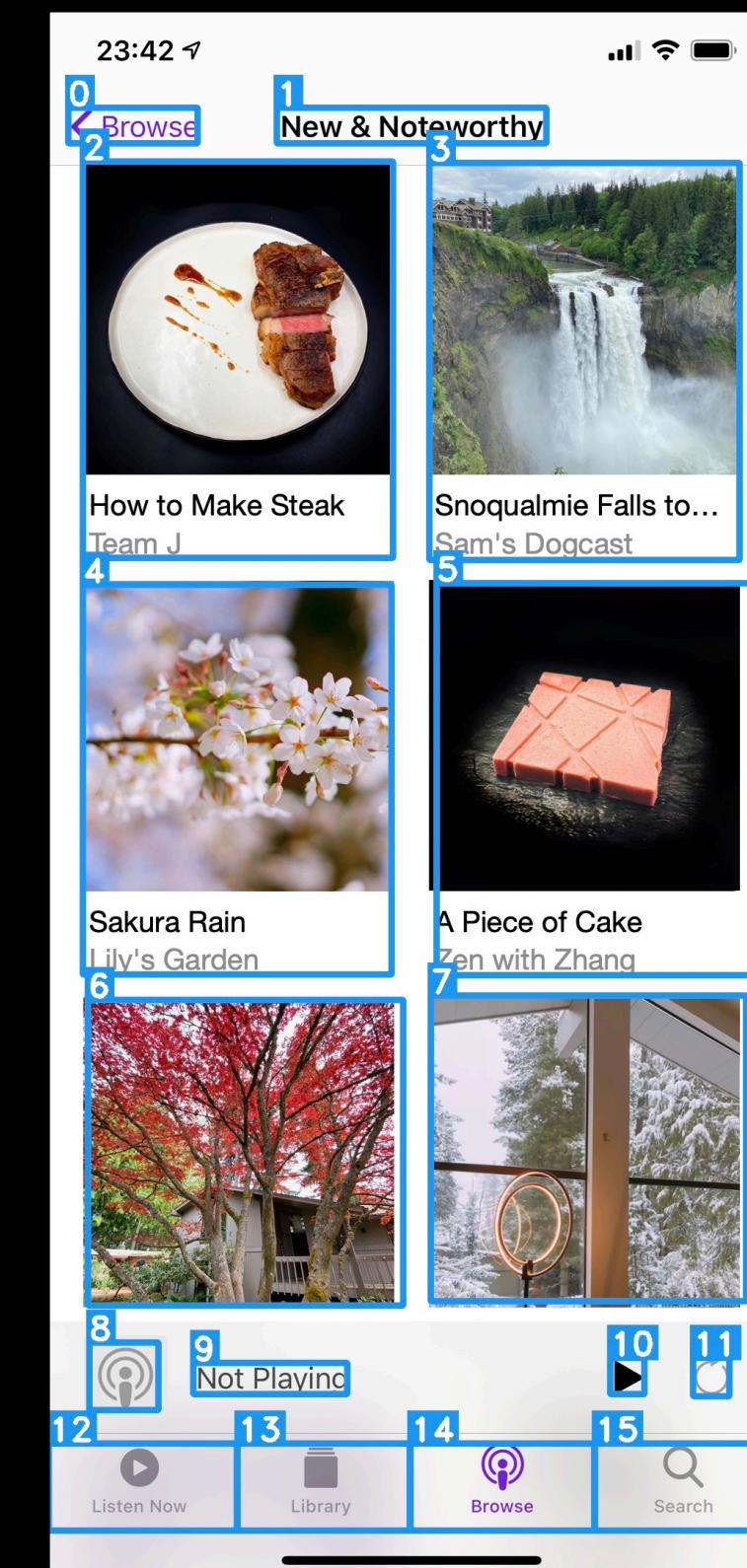
2) Data Annotation



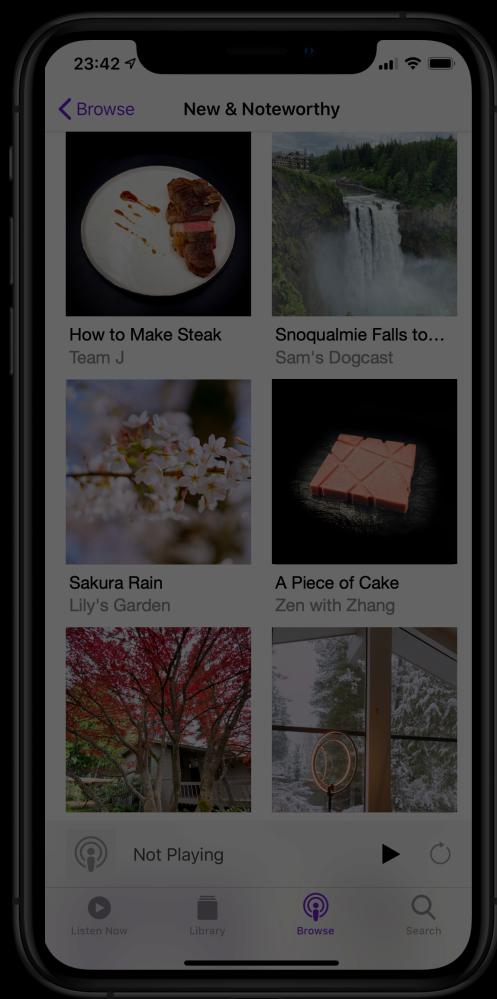
3) UI Detection Model



4) Improved User Experience



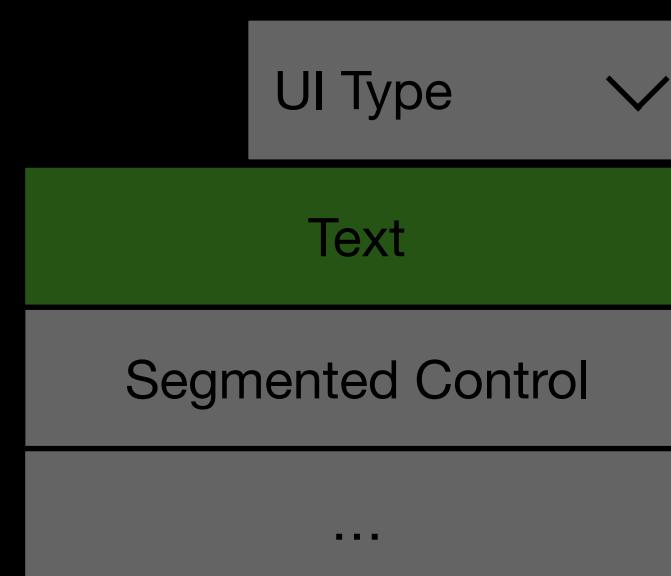
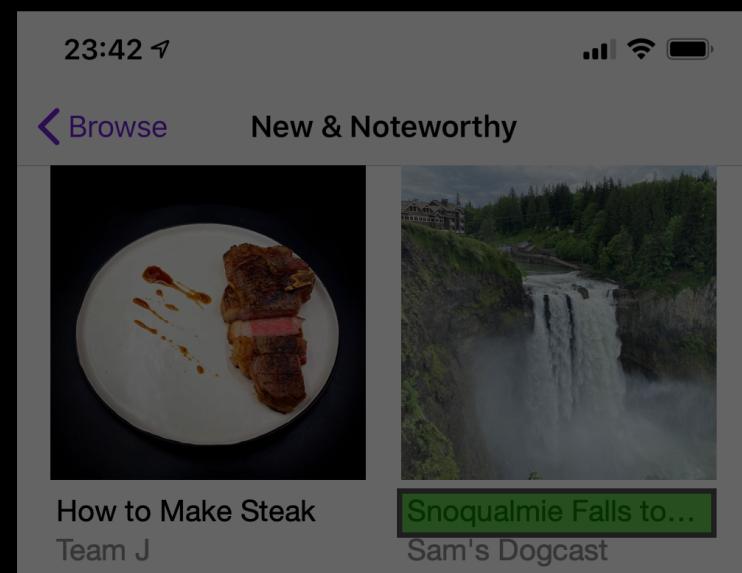
1) Data Collection



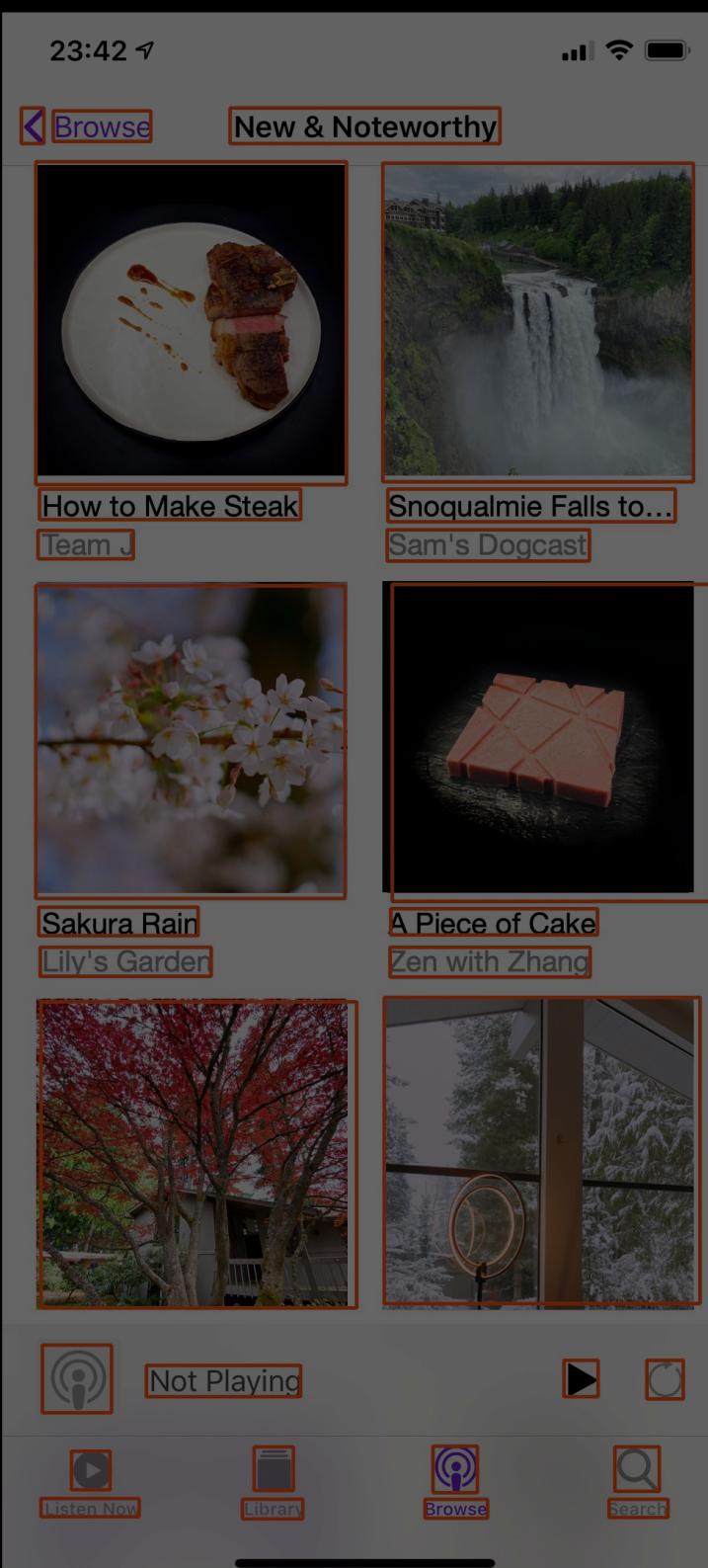
```
"hierarchy": {  
  "accessibilityElements": [ ...  
    {"class": "UIButton",  
     "frame": [10,10, 150, 210]  
    ]  
  ...  
}
```

UI Tree

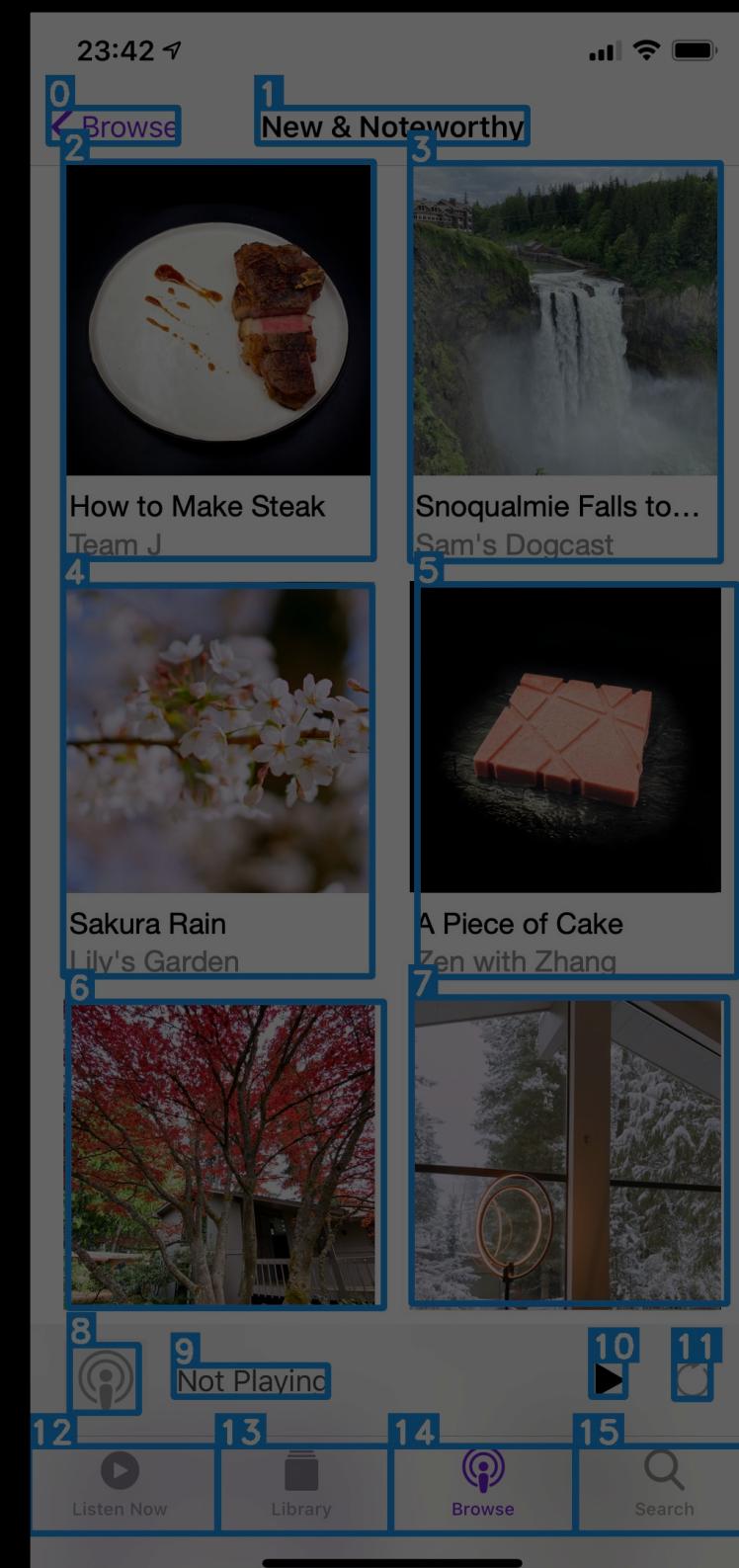
2) Data Annotation



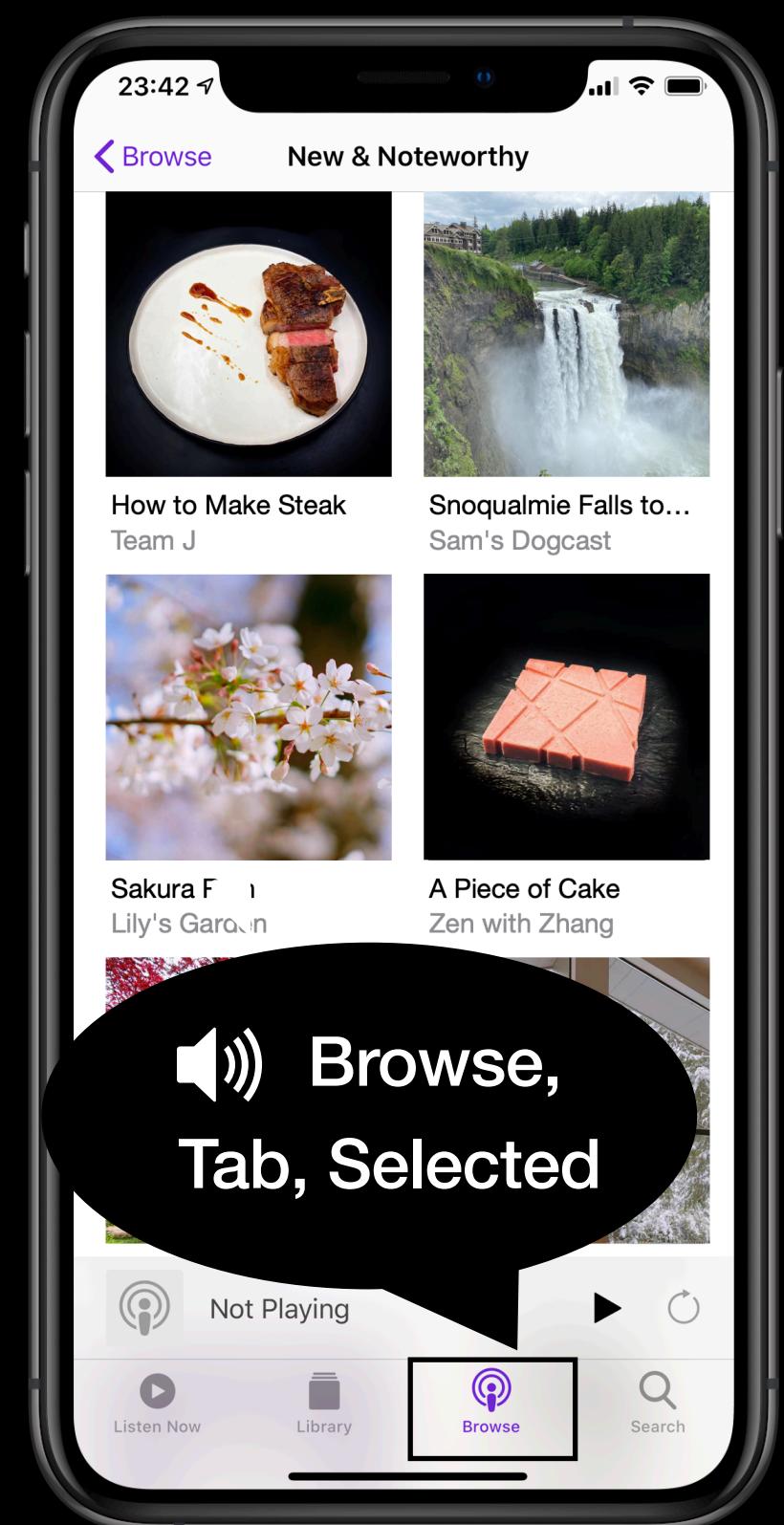
3) UI Detection Model



4) Improved User Experience



5) Integration with Screen Reader



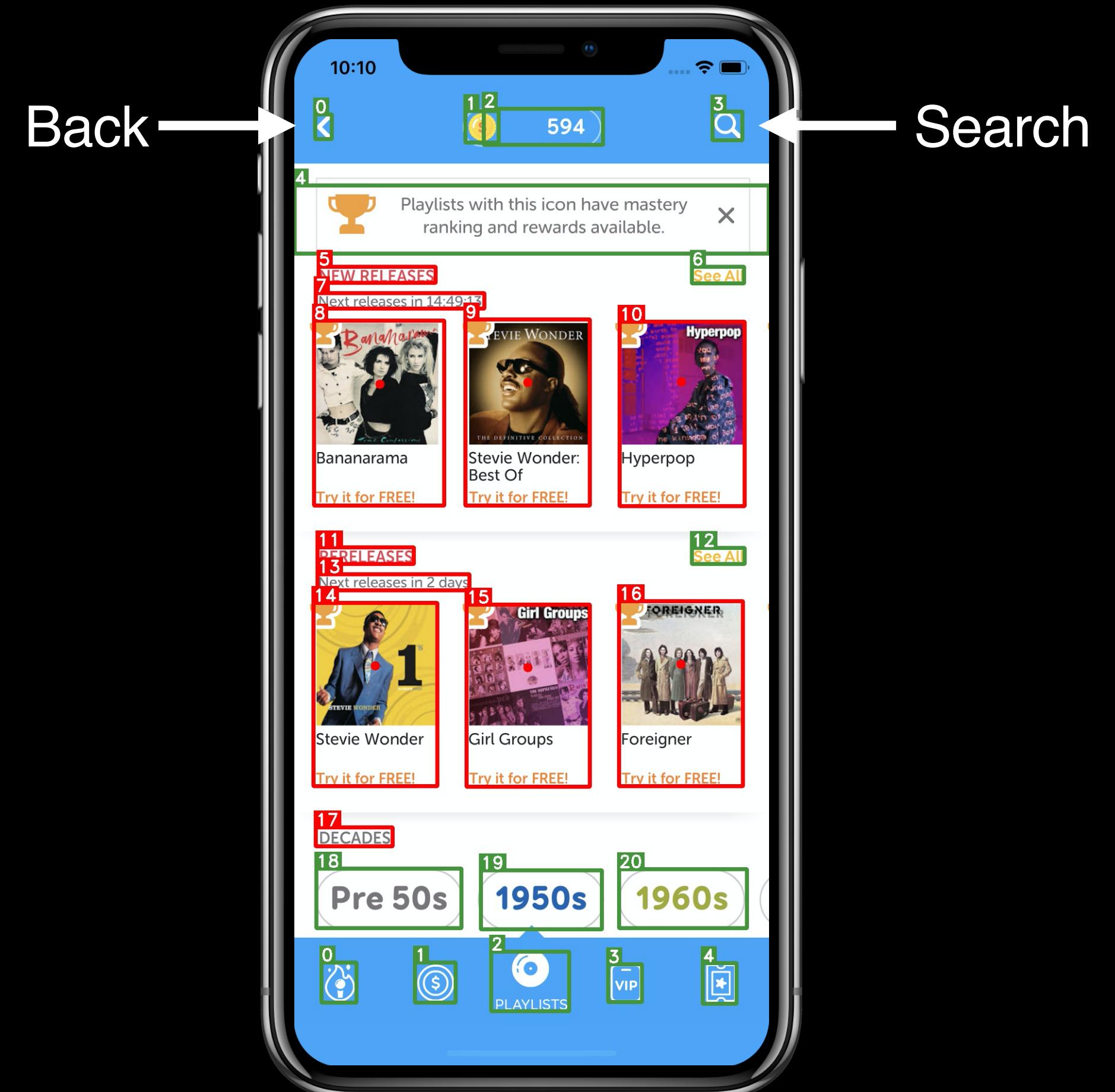
Demo

Towards complete icon labeling in mobile applications

1. Jieshan Chen, Amanda Swearngin, Jason Wu, Titus Barik, Jeffrey Nichols, and Xiaoyi Zhang. "Towards complete icon labeling in mobile applications." In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, pp. 1-14. 2022.

Understanding Icons

- Icons are the second most frequent UI element type (after Text) in mobile apps.
- Icons are often missing accessibility labels.
 - Over two-thirds of image-based buttons are missing labels across 77% of 10,408 mobile apps.

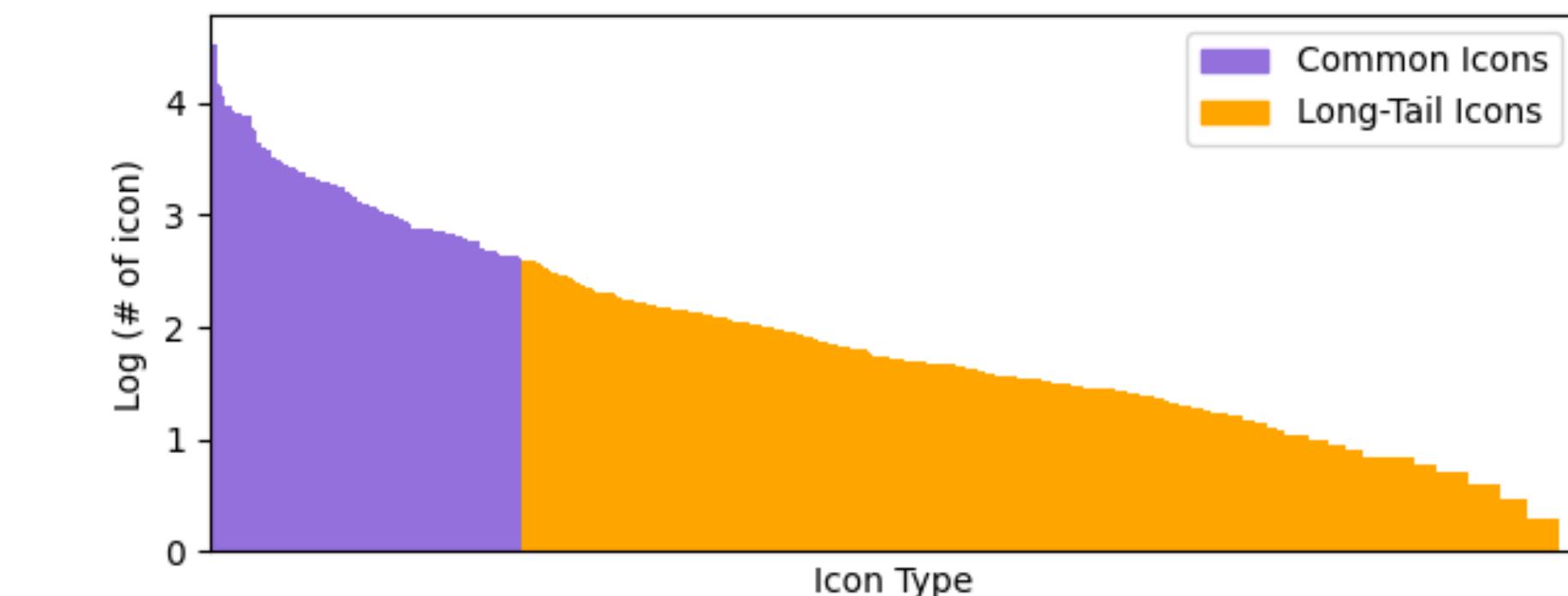
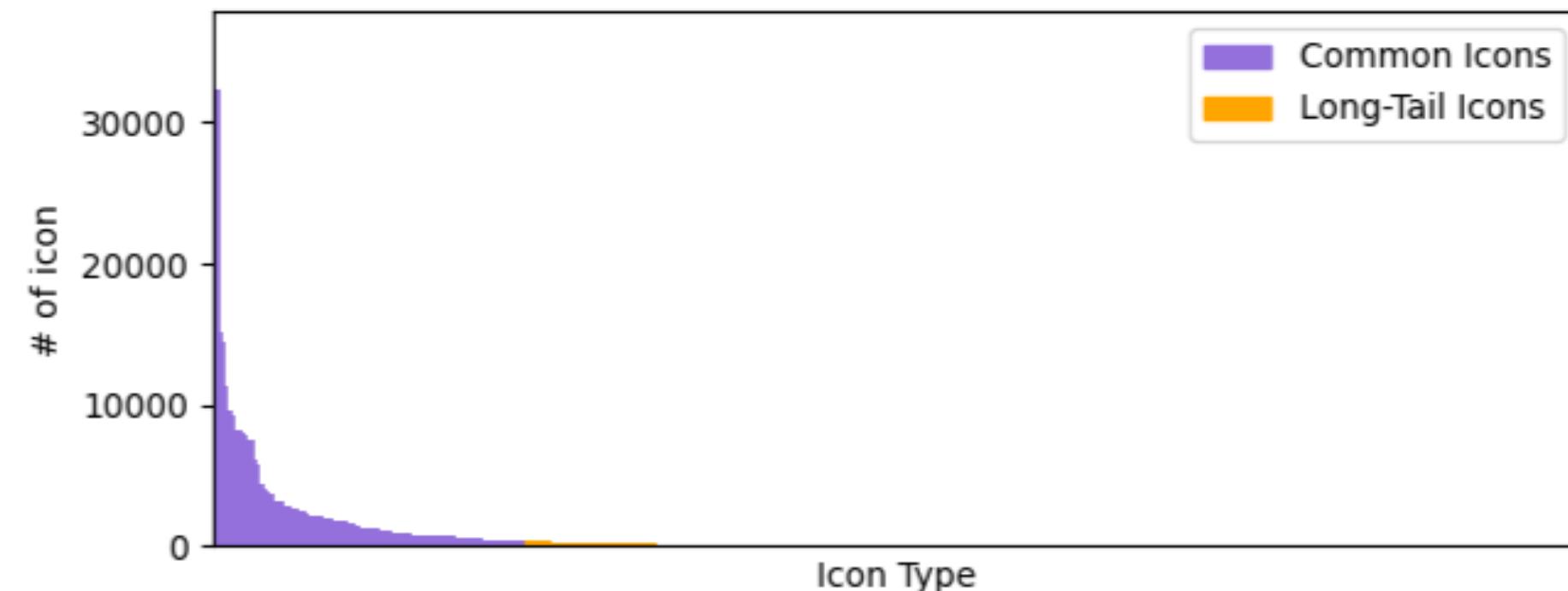


Data Analysis

Icon type distribution

- 98 common icon types + 331 long-tail icon types
- Highly uneven distribution across icon types resembles a long-tail distribution.

| | Common Icons | Long-tail Icons | Unclassified Icons | Total |
|---------------------|--------------|-----------------|--------------------|---------|
| # of Icon Types | 98 | 331 | - | 429 |
| # of Icons | 304,310 | 22,088 | 1,481 | 327,879 |
| # of Screens | 65,906 | 14,376 | 1,234 | 66,138 |
| # of Apps | 3,899 | 2,399 | 739 | 3,904 |
| # of App Categories | 22 | 22 | 22 | 22 |

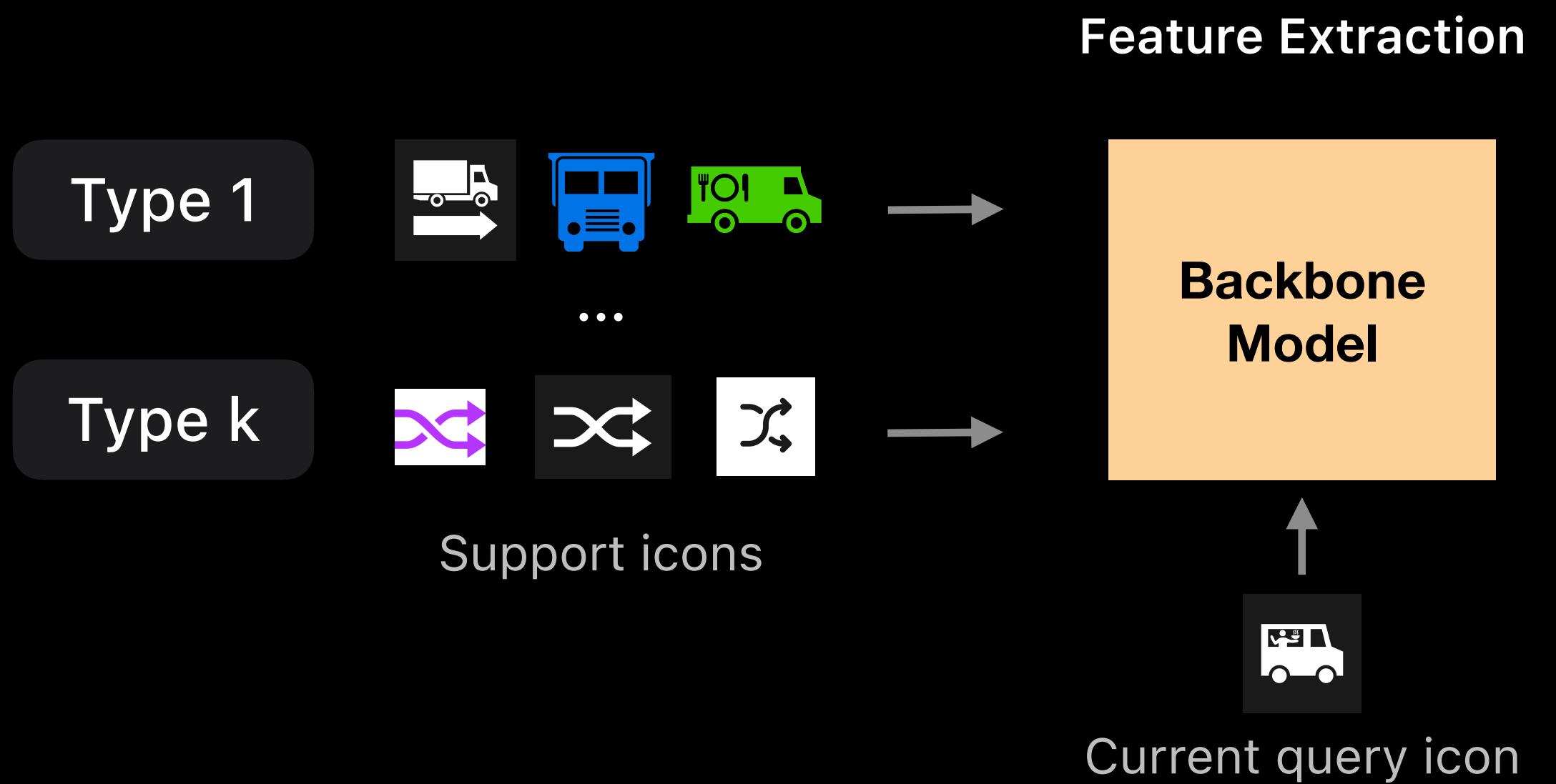


Long-Tail Icon Model

Prototypical network

Backbone Model:

Reuse common icon model, replace the last fc layer, and fine-tune last two layers



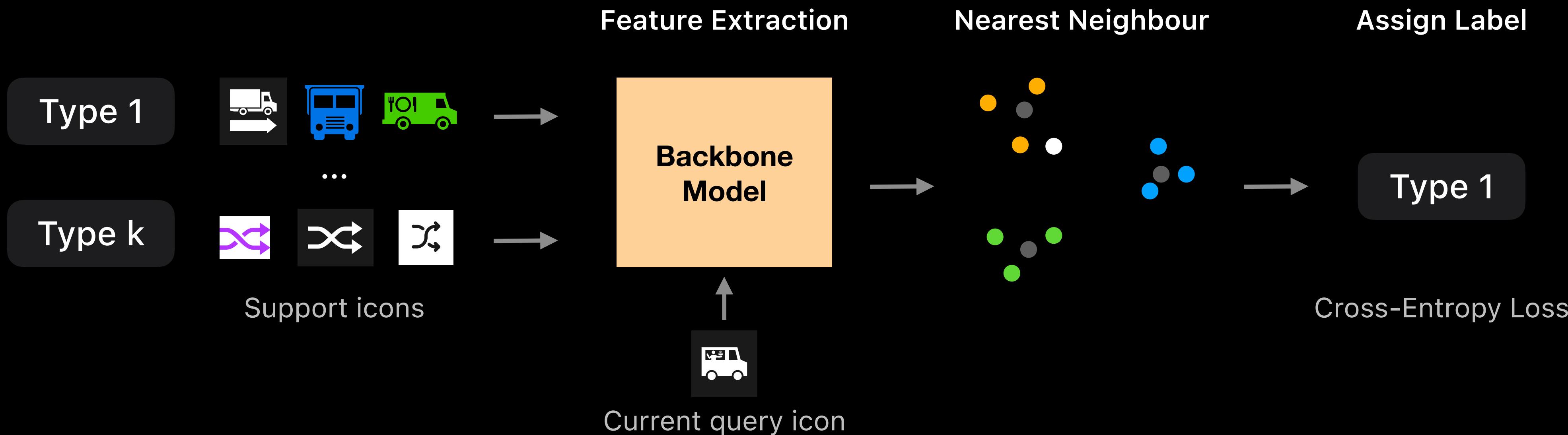
- ● ● Support icons of different icon types
- Mean feature of all support icons of a certain icon type
- Current query icon

Long-Tail Icon Model

Prototypical network

Backbone Model:

Reuse common icon model, replace the last fc layer, and fine-tune last two layers



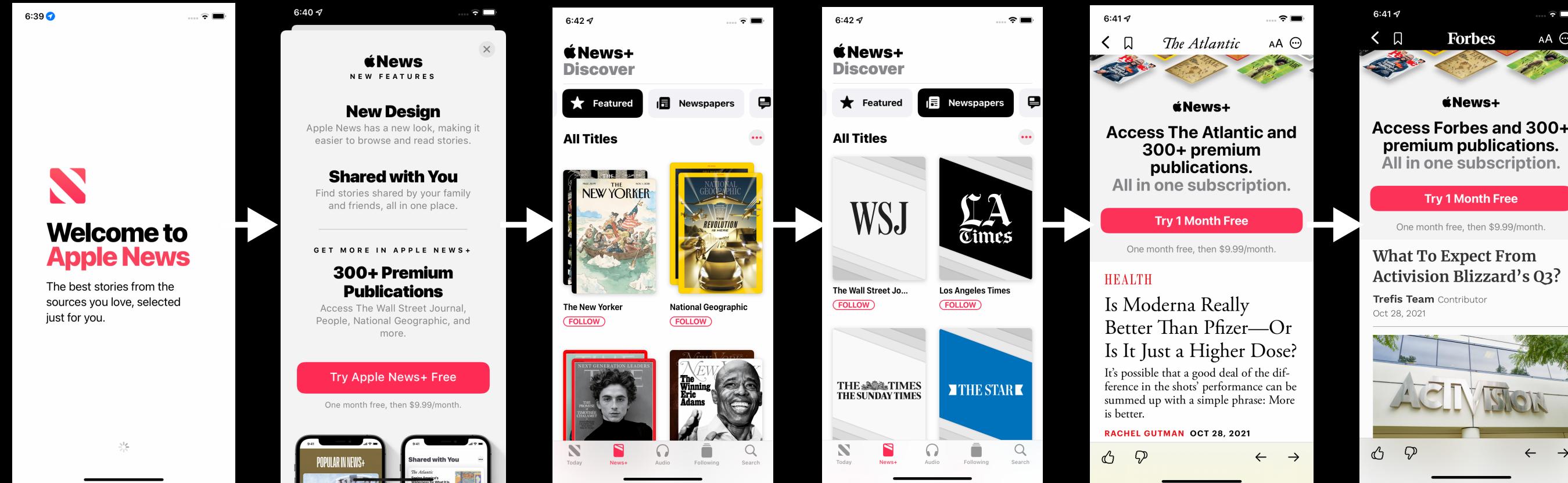
- ● ● Support icons of different icon types
- Mean feature of all support icons of a certain icon type
- Current query icon

Understanding screen relationships from screenshots of smartphone applications

1. Shirin Feiz, Jason Wu, Xiaoyi Zhang, Amanda Swearngin, Titus Barik, and Jeffrey Nichols. "Understanding screen relationships from screenshots of smartphone applications." In 27th International Conference on Intelligent User Interfaces, pp. 447-458. 2022.

Understanding Screen Relationships

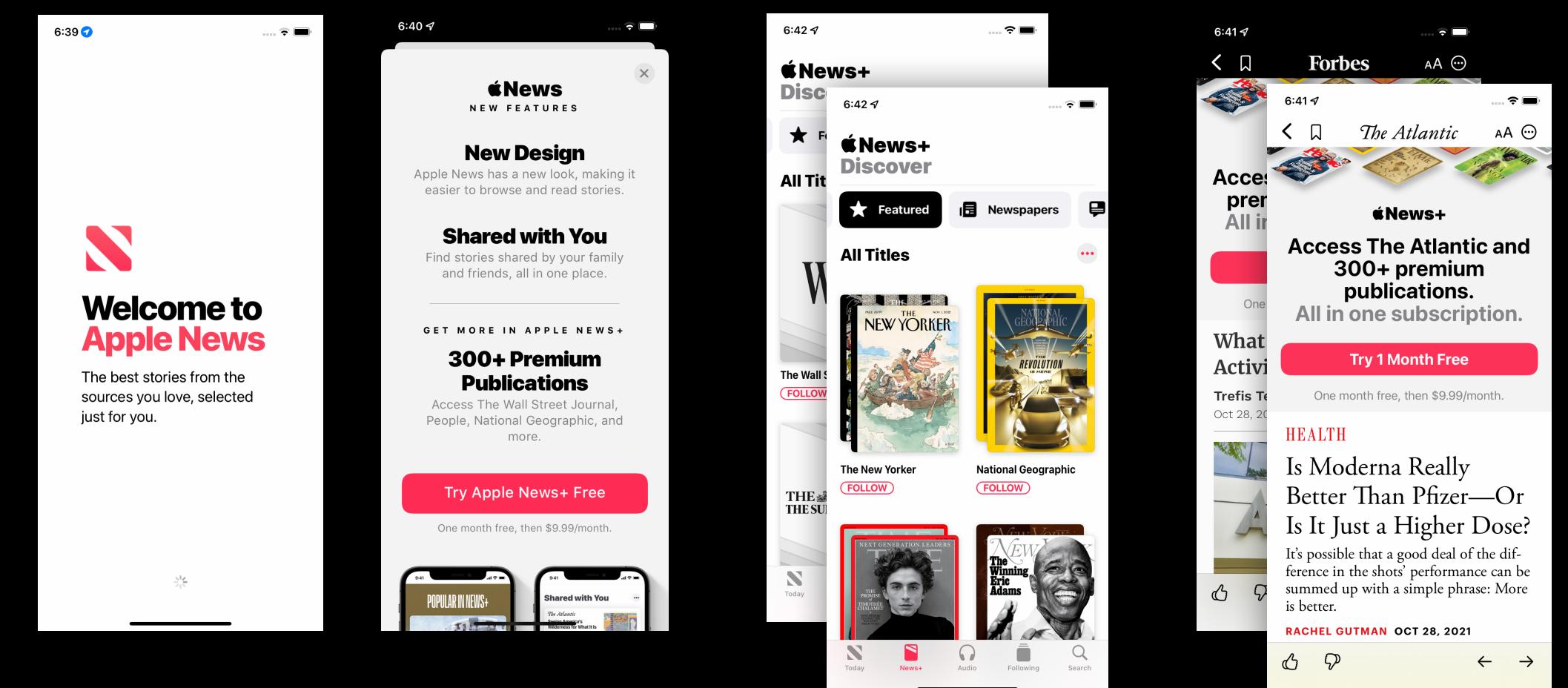
App Crawling: Is the screen the crawler is at now the same as any screen we have seen previously?



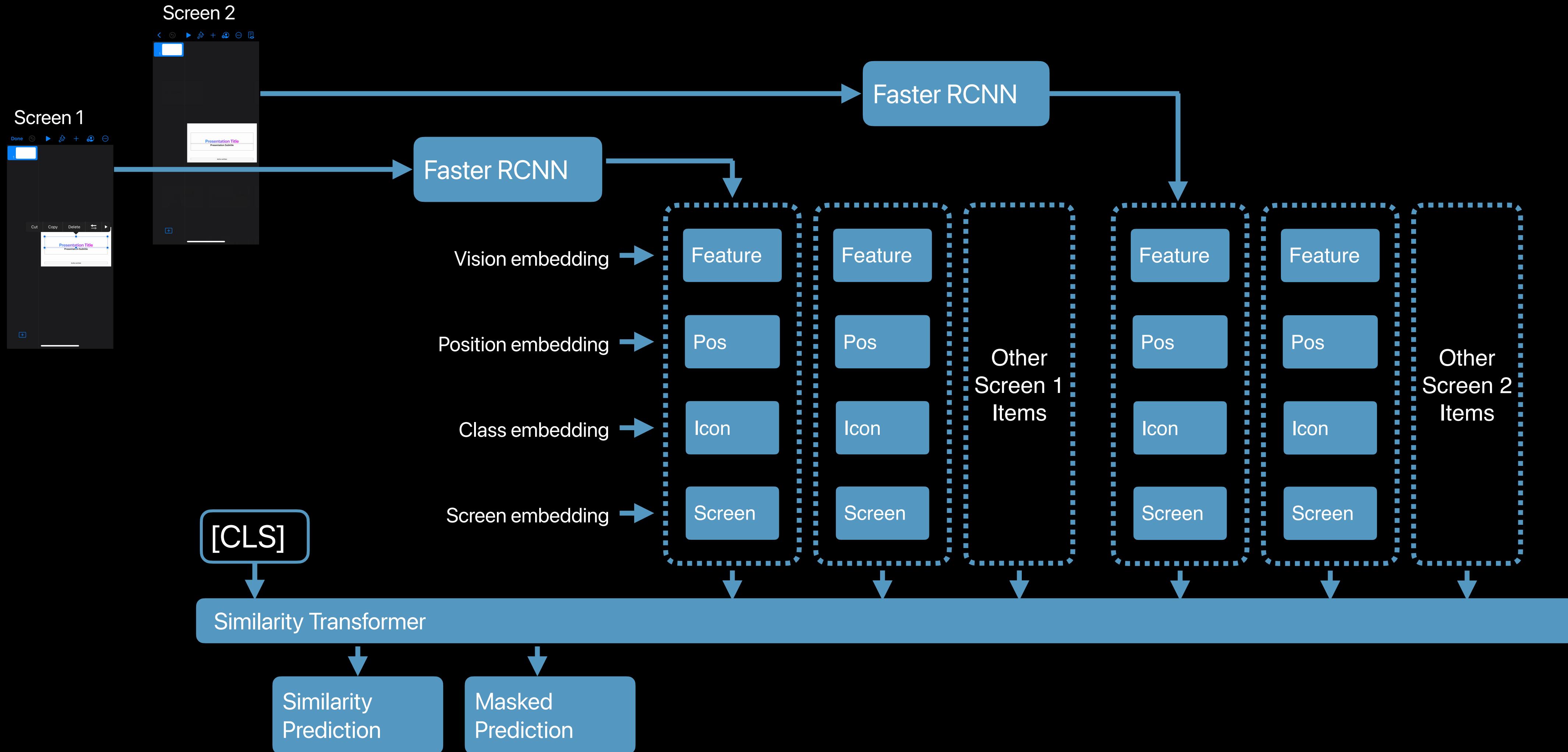
Interaction
Trace

App Crawling: UI
Testing and app
review, accessibility

Grouped
Screens



Screen Similarity: Model Architecture



Screen Similarity: Use cases

App Crawling



Data

Testing

Review

Accessibility Reports

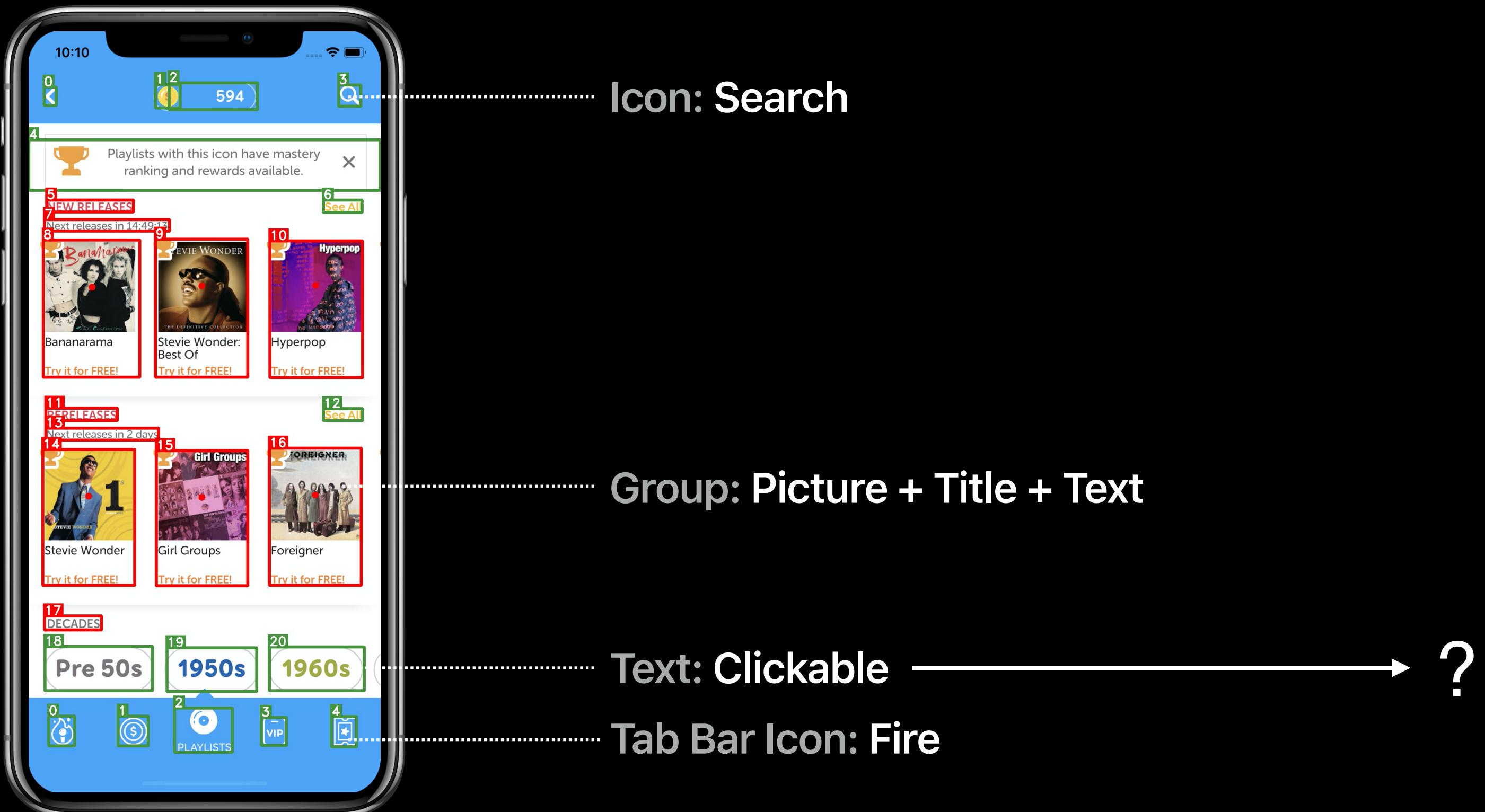
A screenshot of a Mac desktop showing the Accessibility Inspector window and a web browser displaying the Xcode Accessibility Audit Results. The Accessibility Inspector window shows a list of findings for "Screen 5", including issues like "Hit area is too small" and "Dynamic Type font sizes are unsupported". The Xcode window shows a "Summary" section with multiple screenshots of an app interface, followed by sections for "Hit Region", "Element Description", and "Dynamic Type", each listing specific accessibility warnings with counts like "2 warnings", "49 warnings", "17 warnings", "32 warnings", "40 warnings", and "Total Warnings: 91".

Xcode Feature!

Never-ending Learning of User Interfaces¹

1. Jason Wu, Rebecca Krosnick, Eldon Schoop, Amanda Swearngin, Jeffrey P. Bigham, and Jeffrey Nichols. "Never-ending Learning of User Interfaces." In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, pp. 1-13. 2023.

Limitations of static representation

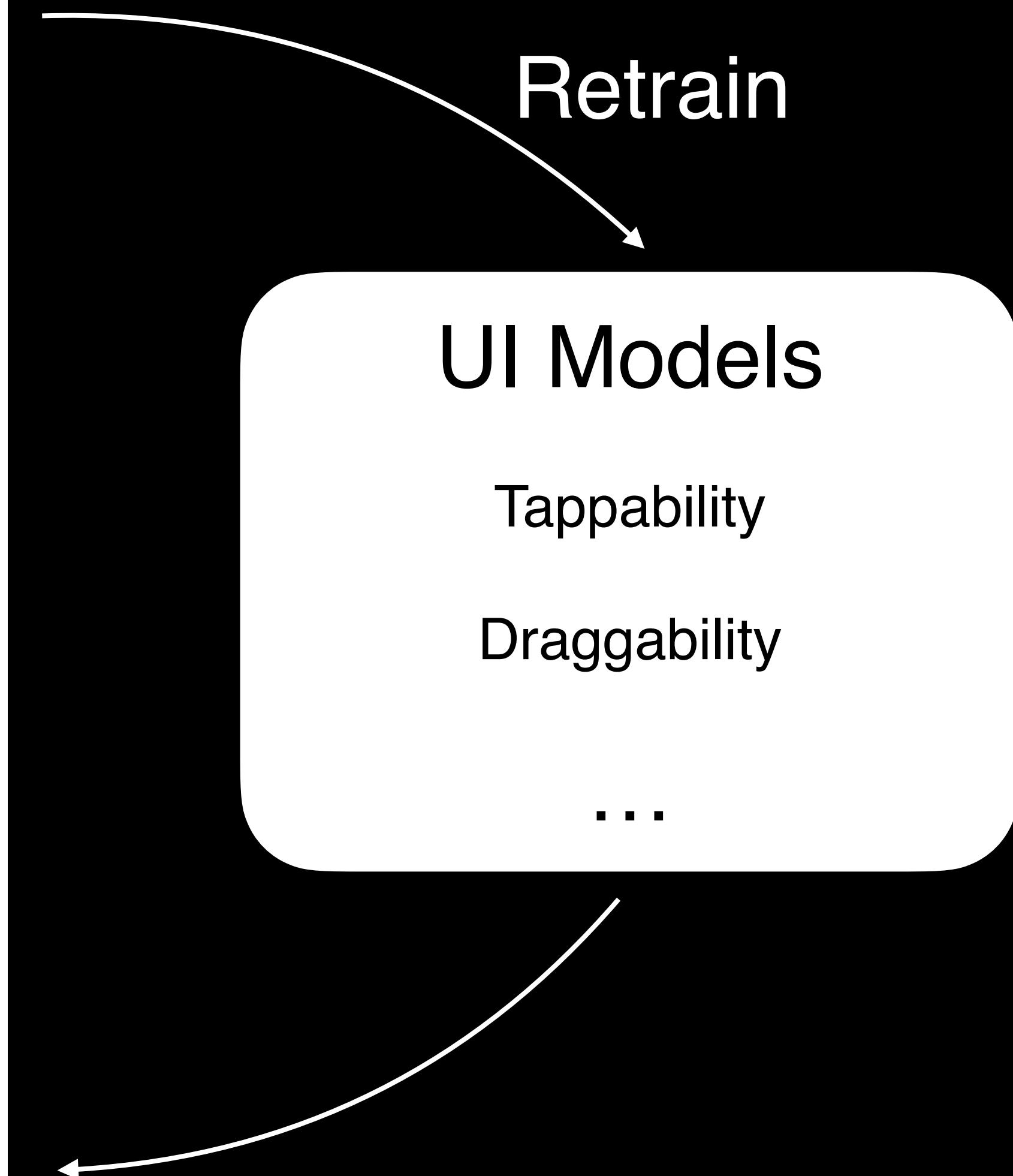
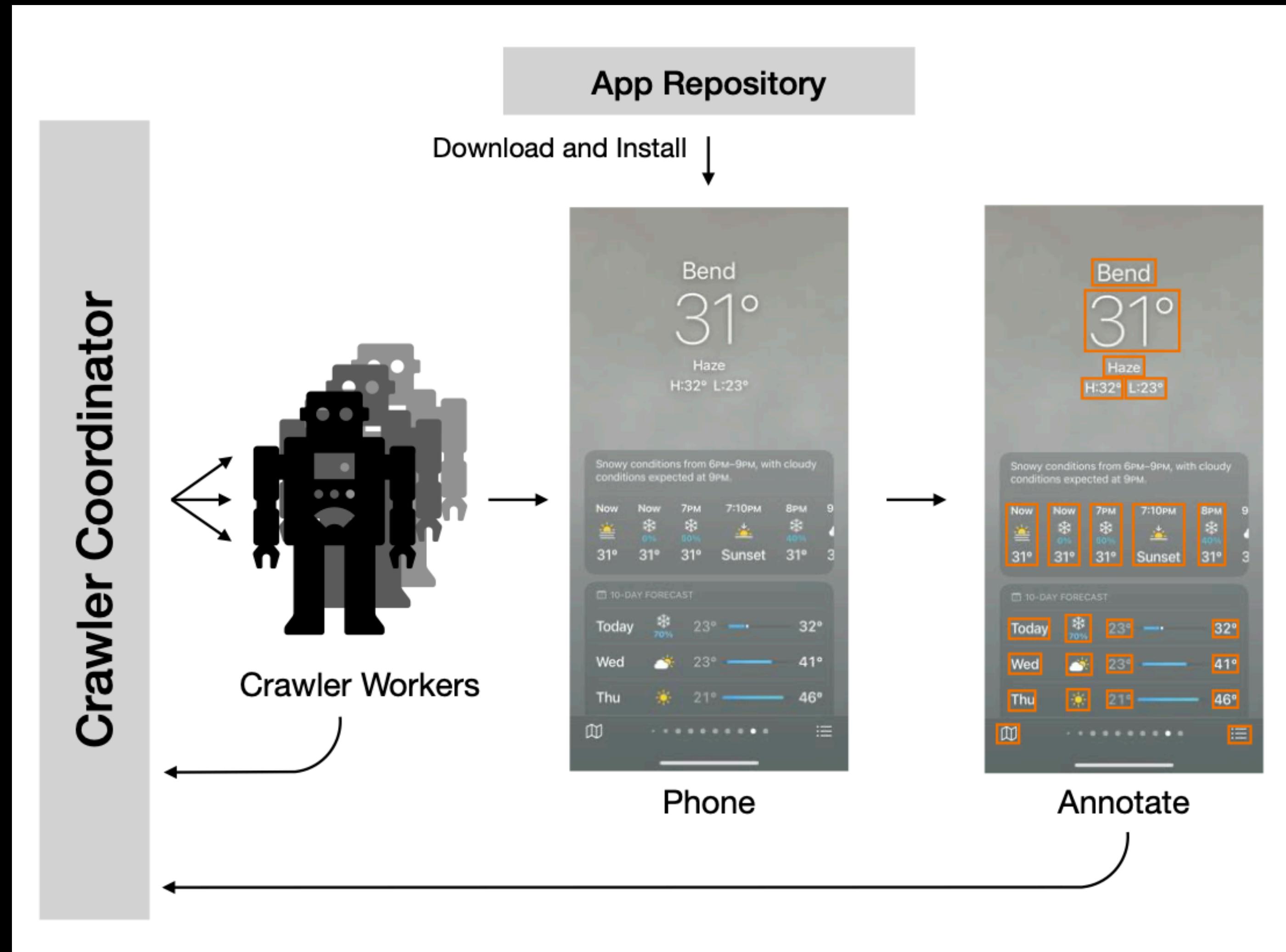


Icon: Search

Group: Picture + Title + Text

Text: Clickable → ?

Tab Bar Icon: Fire



Questions?

Training a screen recognizer Using PyTorch + Lightning

Exercise 1

Objectives:

- Train UI element recognition model
- Evaluate its output
- Improve the output for downstream cases (screen reader, LLM, etc.)

Bonus Objectives

Training a screen classifier

Training a screen similarity model

VINS Dataset

Annotated UI dataset

- Abstract wireframes
- High fidelity UIs
- annotated with bounding boxes and UI types

4,543 screens

- iPhone, Android

WebUI Dataset

WebUI: A Dataset for Enhancing Visual UI Understanding with Web Semantics¹

- 400K web UIs
- Collected from automated web crawling
- Goal: Support transfer learning of UI data
- Huge thanks to **Jason Wu** for contributing his dataset and starter code for the screen recognition exercises!
- <https://uimodeling.github.io/>



1. Wu, Jason, et al. "Webui: A dataset for enhancing visual ui understanding with web semantics." Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 2023.

Part 2

**UI Understanding + LLMs +
VLMS**

Examples

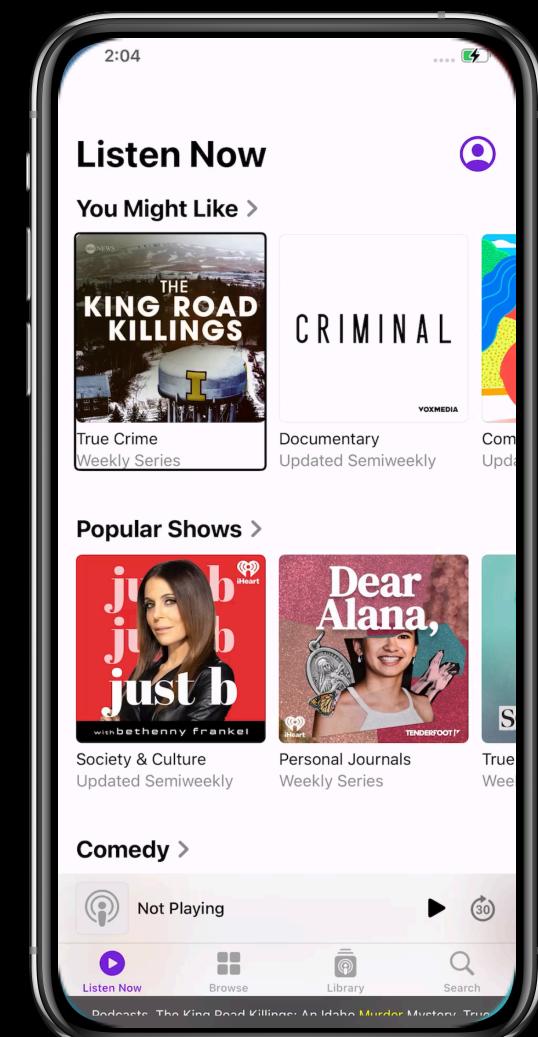
AXNav: Replying accessibility tests from natural language¹

1. Taeb, Maryam, Amanda Swearngin, Eldon Schoop, Ruijia Cheng, Yue Jiang, and Jeffrey Nichols. "Axnav: Replying accessibility tests from natural language." In Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1-16. 2024.

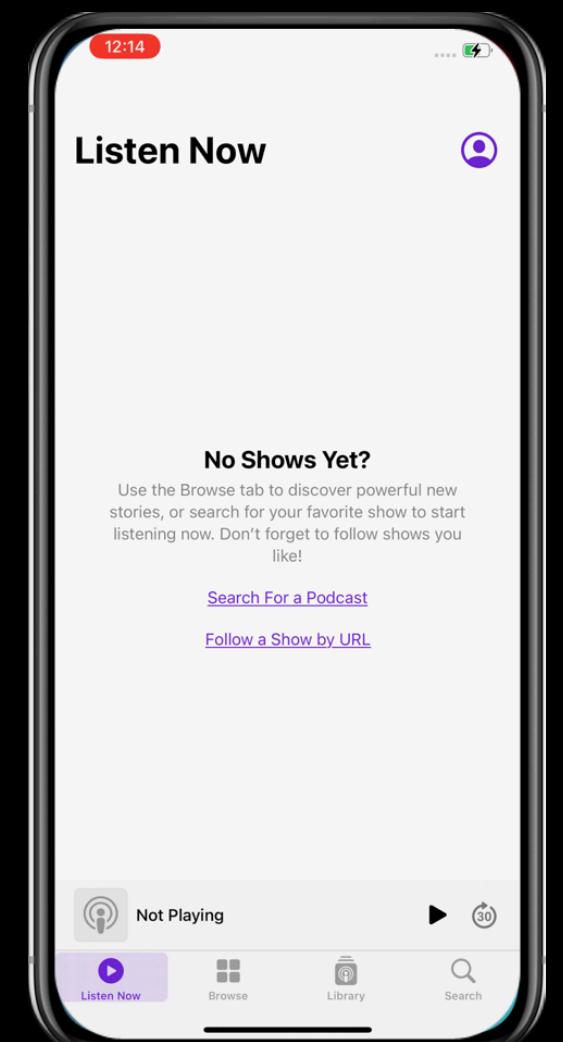
Motivation

Manual Accessibility Testing

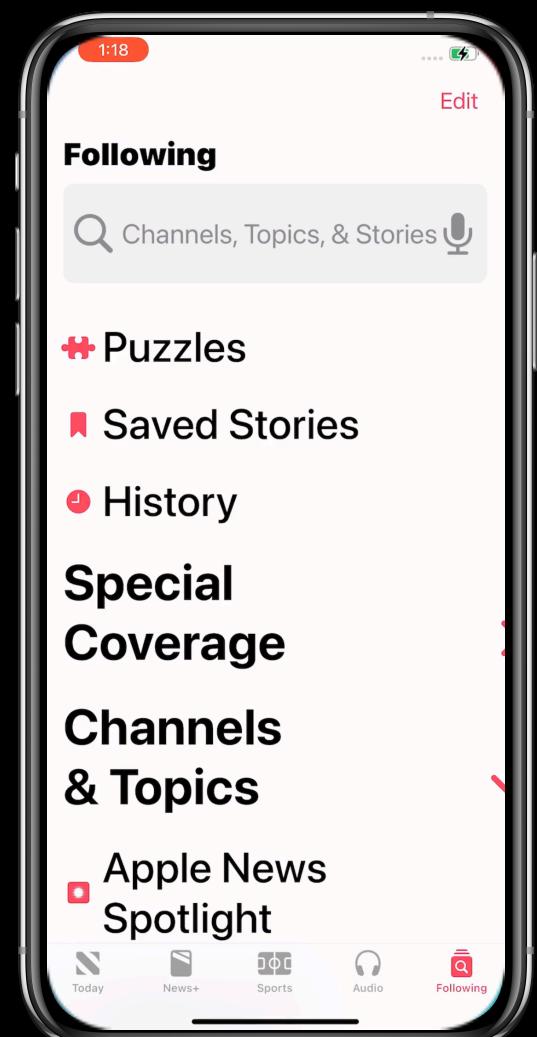
| Title: iOS: VoiceOver: Search for a Show |
|---|
| 1. Go to Settings > Accessibility > VoiceOver, and enable VoiceOver (VO) |
| 2. Launch the Media app |
| 3. Search for a show and verify that everything works as expected and there are accurate labels |
| 4. Turn off VO and verify that searching for a show works as expected |



| Title: iOS: Media App: Dynamic Text in Search Tab |
|--|
| 1. In Settings > Accessibility > Display & Text Size, enable larger text and set to maximum size |
| 2. Launch the Media app |
| 3. Verify all text (titles, headers, etc.) font size has adjusted consistently |
| 4. Set text size to minimum and repeat step 3 |
| 5. Reset text size to default and verify all text returns to normal |



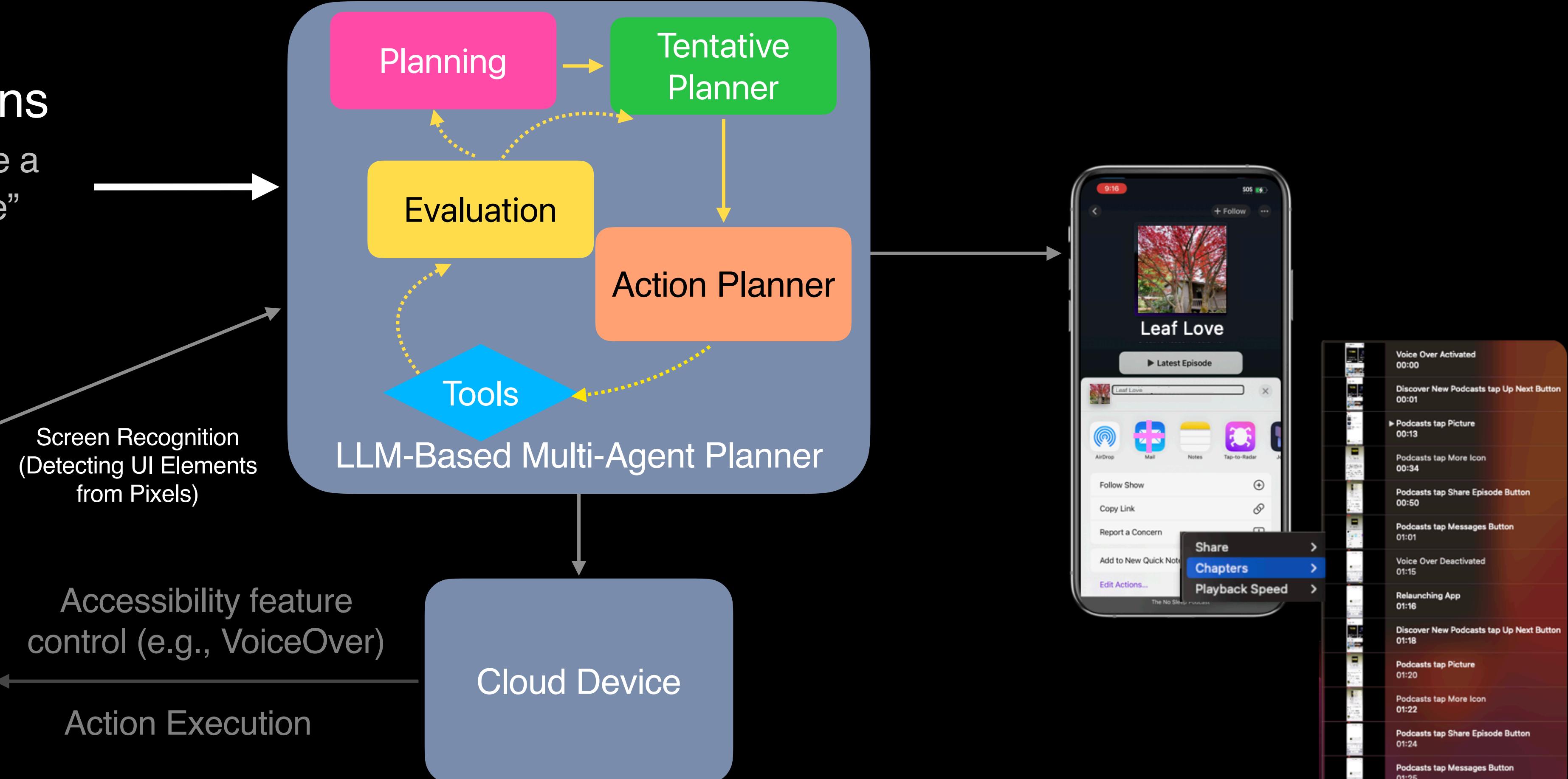
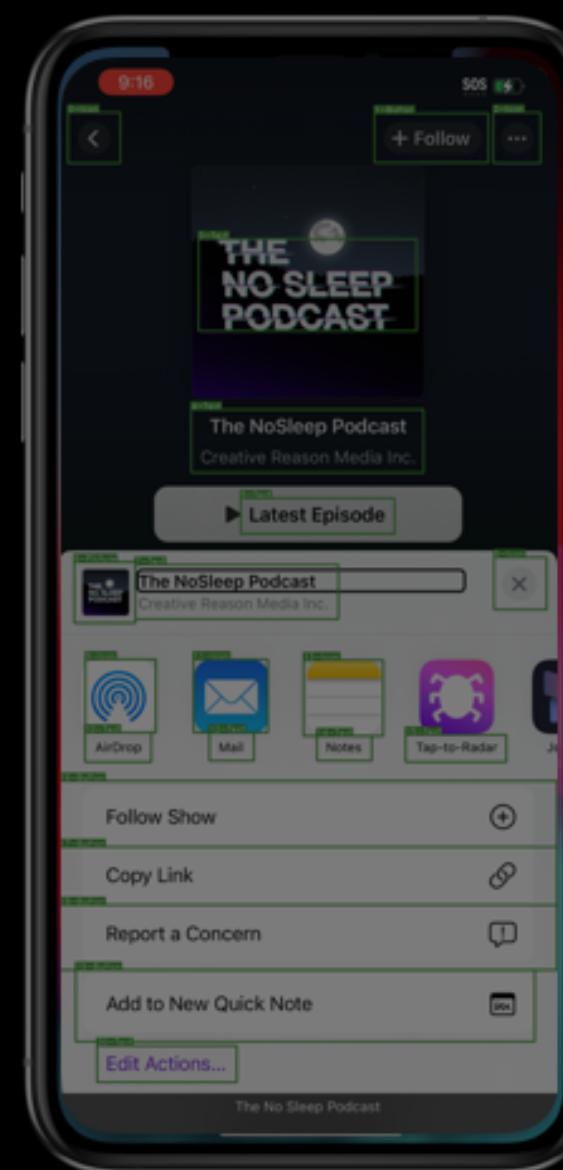
| iOS: Media App: Button Shapes across app |
|---|
| Expected Result: When Testing button shapes- we want to make sure that all text (not emojis or glyphs) get underlined if they are NOT inside of a button shape already. If the text is already within a button shape, it is a bug! (We see this bug frequently) |



AXNav System

Test Instructions

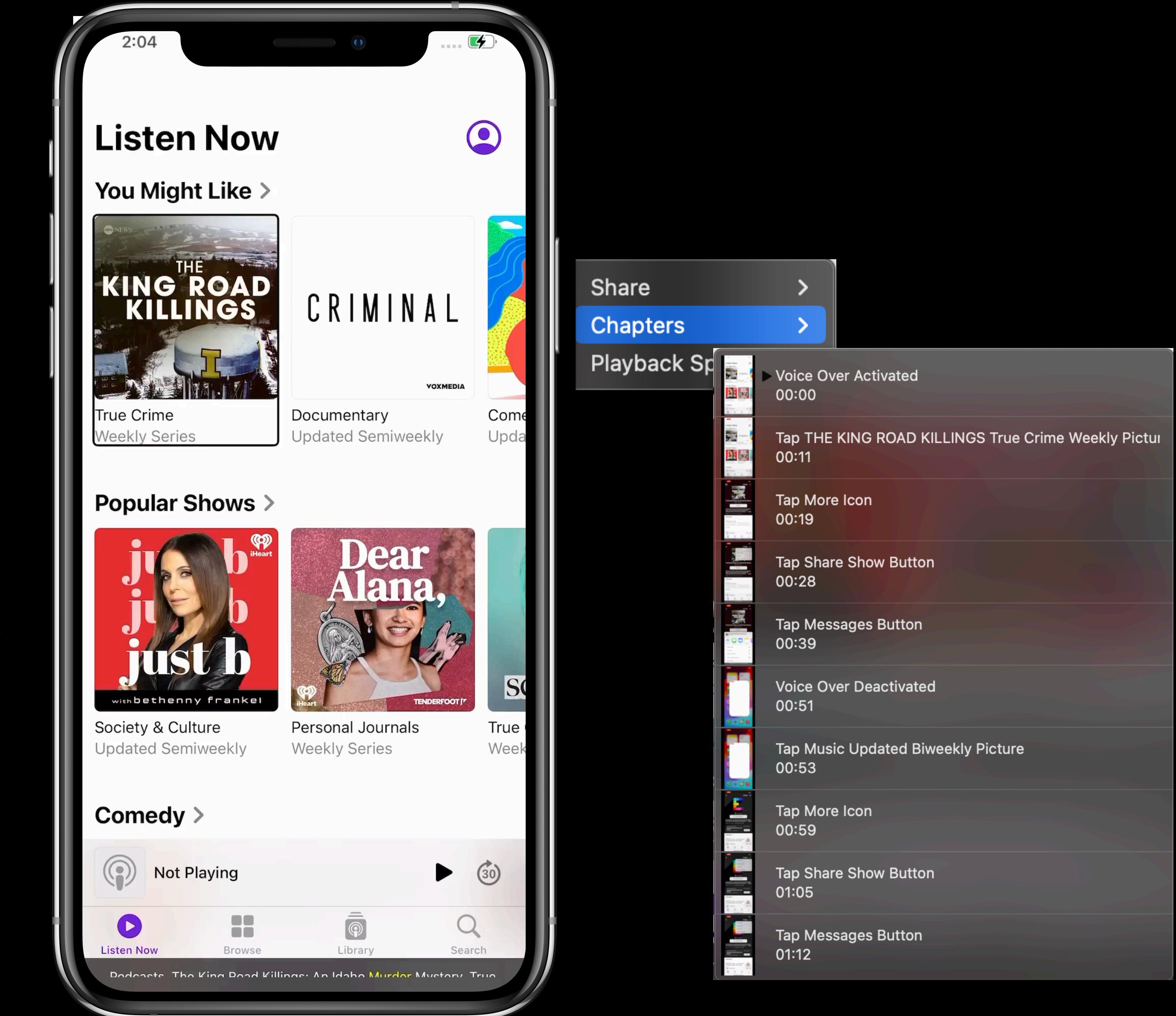
“VoiceOver: Share a Podcast Episode”



AXNav - Demo

Podcasts -
VoiceOver
“Share an Episode”

AXNav

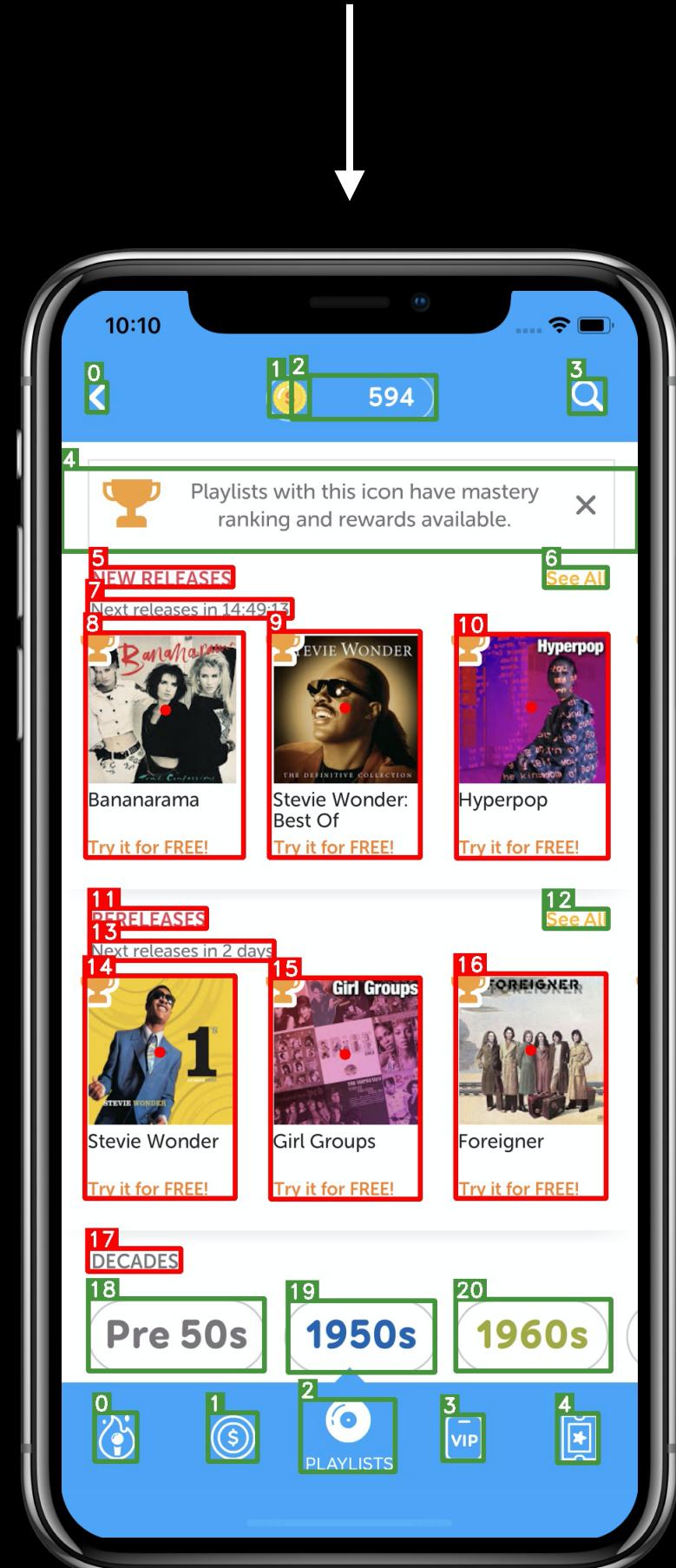


ILuvUI: Instruction-tuned language-vision modeling of uis from machine conversations¹

Ferret-UI: Grounded Mobile UI Understanding with Multimodal LLMs²

1. Jiang, Yue, Eldon Schoop, Amanda Swearngin, and Jeffrey Nichols. "ILuvUI: Instruction-tuned language-vision modeling of UIs from machine conversations." arXiv preprint arXiv:2310.04869 (2023).
2. You, Keen, Haotian Zhang, Eldon Schoop, Floris Weers, Amanda Swearngin, Jeffrey Nichols, Yinfai Yang, and Zhe Gan. "Ferret-UI: Grounded Mobile UI Understanding with Multimodal LLMs." arXiv preprint arXiv:2404.05719 (2024).

Multimodal large language model (VLM)



Lack of training data on UIs

Have not been trained on
UI understanding tasks - e.g.

- Screen summarization
- Screen Q&A
- Elementary tasks (widget classification, OCR, etc.)

Ferret-UI

Elementary Tasks

Referring
widget classification, icon recognition, OCR

- Classify the UI type of the widget [box0].
- Button
- Identify the icon type of the widget [box1].
- launch
- What text is displayed in the element [box4]?
- Reminders
Don't forget. Use reminders

Grounding
find widget, find icon, find text, widget listing

- Where is the Open button?
- Open button [box0]
- Where is the launch icon?
- launch icon [box1]
- Where is the text "Reminders
Don't forget. Use reminders"?
- Reminders, Don't forget. Use reminders [box4]
- Please list all the UI elements on this screen.

UI widgets present in this screen include text displaying "Reminders, Don't forget. Use reminders" [box4], Open Button [box0], Launch icon [box1], ...

Advanced Tasks

Conversation (perception + interaction)

- How do I open the Reminders app?
You can open the Reminders app by tapping on the Open button [box0].
- I want to share the Reminders app with my friend, what should I do?
You can share the app by tapping on the launch icon [box1].
- How do I navigate to Games?
You can navigate to the Games tab by tapping on the Games tab [box2].
- Can my 15-year-old daughter use this app?
Yes, the app is suitable for all ages as indicated by the "4+" age rating [box3].

Detailed Description

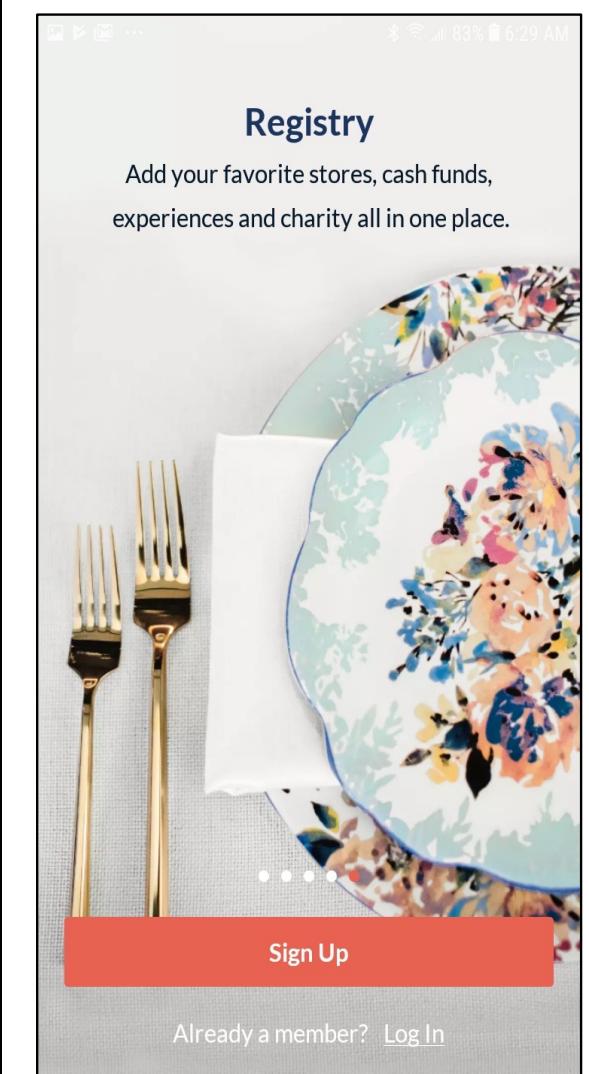
Please describe this screen in detail.

The screen is for a reminders application. At the top left, there's a button labeled "Apple". The title of the screen, "Reminders", is located in the upper middle part of the screen. There's a text below the title that reads "Don't forget. Use reminders".

The screen is divided into several sections. The first section, labeled "210k+", is located just below the title. The second section, labeled "AGE", contains a text displaying "4+ Years Old" ...

Function Inference

What's the high-level function of the screen?
The screen is the download page for a reminders application where you can set up reminders for various tasks such as sending out team's weekly progress, grocery shopping, traveling, and picking up kids.



UI Element Existence:

Positive Sample:

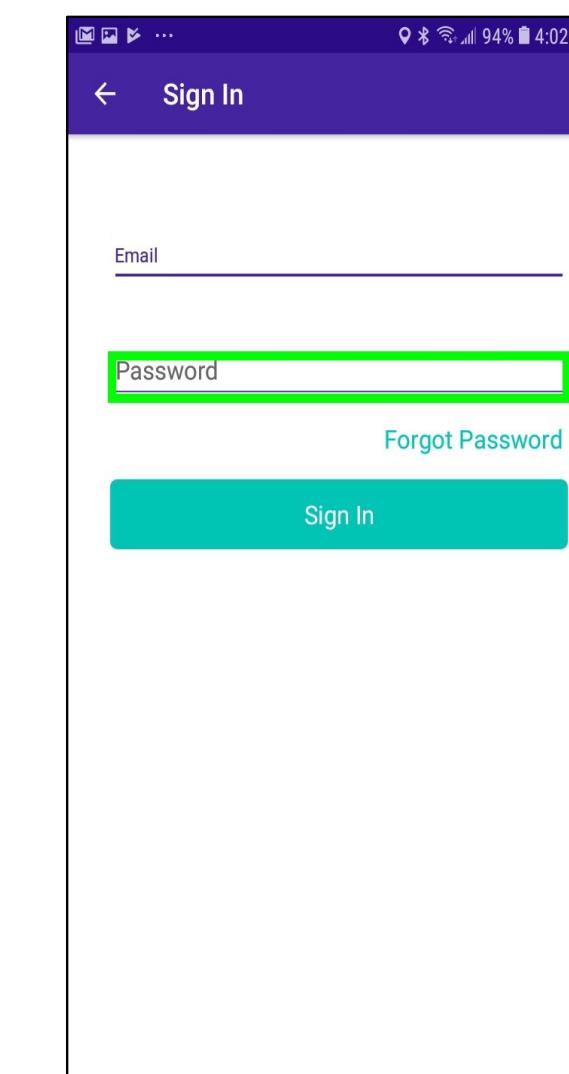
Human: Does the "Sign Up" Text Button exist in the UI?

ILuvUI: Yes.

Negative Sample:

Human: Does the "back" Icon exist in the UI?

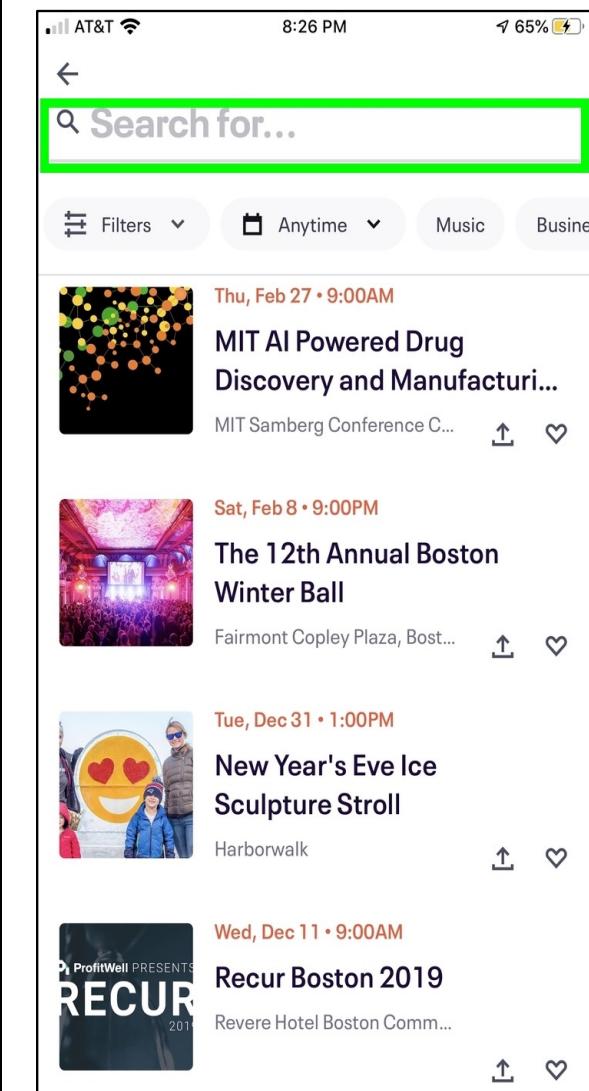
ILuvUI: No.



UI Element Purpose:

Human: What is the purpose of the UI element in the green bounding box?

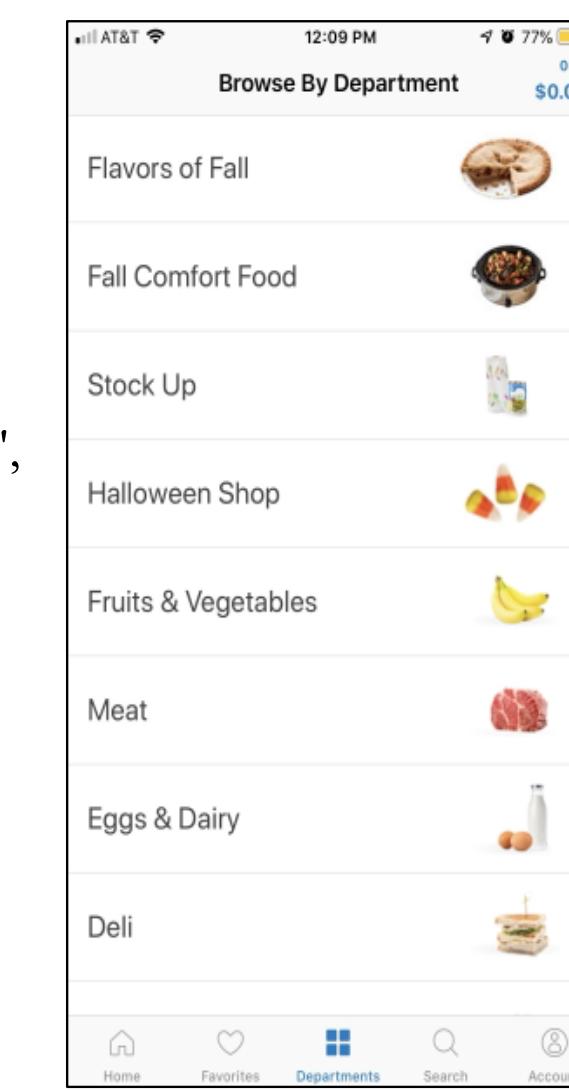
ILuvUI: It is a text field where users can enter their password for signing in.



UI Element Type:

Human: What is the type of the element in the green bounding box? (select from [Icon, Input, Radio Button, Toolbar, Switch, Check Box, Checked Text View, Upper Task Bar, Text Button, Text Field, Page Indicator, Image, Text])

ILuvUI: Text Field.



Summarization:

Human: Provide a one-sentence summarization.

ILuvUI: This UI screen displays a browse-by-department feature with various categories such as bread & bakery, pantry, snacks & candy, beverages, frozen, organic shop, special dietary needs, and household essentials, along with tabs for home, favorites, departments, search, and account.

Evaluate using LLMs and VLMs for UI Q&A

Exercise 2

Objectives

Use notebook to query LLM

Provide UI detection info to the LLM

Develop and evaluate your own screen Q&A